SWITCHING LINEAR DYNAMIC SYSTEMS WITH HIGHER-ORDER TEMPORAL STRUCTURE

A Dissertation
Presented to
The Academic Faculty

by

Sang Min Oh

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
College of Computing

Georgia Institute of Technology
August 2009

SWITCHING LINEAR DYNAMIC SYSTEMS WITH HIGHER-ORDER TEMPORAL STRUCTURE

Approved by:

Frank Dellaert, Advisor
College of Computing
Georgia Institute of Technology

Irfan Essa
College of Computing
Georgia Institute of Technology

James M. Rehg
College of Computing
Georgia Institute of Technology

Padhraic Smyth
Dept. of Computer Science
Univ. of California, Irvine

Aaron Bobick
College of Computing
Georgia Institute of Technology

Date Approved: 2009

To my Family :

Yeeun & Jahin

Mother & Father

ACKNOWLEDGEMENTS

During my Ph.D programs, I have been greatly privileged to work with two exceptional academic advisors, Prof. Frank Dellaert and Prof. James M. Rehg. Frank welcomed me into his research group of excellent students in my early years in the program and always has been a great mentor both academically and personally. I owe Frank a large portion of my ways to learn, think, communicate, and be good, which I gladly plan to attain through the rest of my life. When I started to delve into SLDSs, which finally is the topic of this dissertation, Jim helped me to lay ground on the problems with his expertise and has stimulated me to be a constant meta-thinker, a person who thinks about what one thinks. Overall, I am truly grateful to my both advisors for helping me to build my own views which would grow to be broader and deeper over time.

I also thank my thesis committee members, Prof. Aaron Bobick, Prof. Irfan Essa, and Prof. Padhraic Smyth, who shared their valuable time to discuss my work and my future. Their constructive suggestions really helped me to improve this dissertation as a more concrete work. Among other faculty members, I would like to thank Prof. Tucker Balch, and Prof. Bruce Walker, with whom I had chances to work with and sometimes to be even financially supported.

It was all possible to write this dissertation since I was with the bright colleagues. In particular, two of my lab seniors, Michael Kaess and Ananth Ranganathan greatly helped me in all dimensions, ranging from tedious daily happenings to discussion on dissertation writing, to whom I owe a lot. I am also greatly thankful to all the lab-mates and visitors in BORG, Wall, and CPL lab, and members in SSH and KSA, with whom I shared ups and downs in work and life through the years. In particular, I shared fun outside of GeorgiaTech with Demba Ba, Sasa Junuzovic, Hosung Kang and Juyong Kim. I could share Korean warmth with Dongshin Kim, Kihwan Kim, Jungsoo Kim and Junseok Shin. Grant Schindler and George Baah shared the adventures of Ph.D program together, and David Minnen and

Jeff Wilson helped my work substantially in many ways.

During my program, I was lucky enough to deviate to other research environments where I could broaden my horizons. I would like to thank Dr. Cha Zhang, Dr. Paul viola, and Dr. Nebojsa Jojic for hiring me as interns and giving me generous freedom to work on really interesting problems.

In addition, I would like to thank Samsung Lee Kun Hee scholarship foundation for supporting me in the first four years of my program. Their financial support allowed me to work creatively, and the bright young people I met at their annual meetings stimulated me intellectually.

Most of all, I was able to finish this dissertation since I have been given the greatest and deepest support from my family. My father has always encouraged me to be professional and good in my doings and my mother sent me endless mental support during my program which now dates thirty years. Furthermore, I married my wonderful wife Yeeun who supported me with unbelievable love and patience. With Yeeun, I now have my adorable first daughter Jahin, along with the best family-in-laws. My family is the source of my energy and I would dedicate this dissertation to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Automated analysis of temporal data is a task of utmost importance for intelligent machines. For example, ubiquitous computing systems need to understand the intention of humans from the stream of sensory information, and health-care monitoring systems can assist patients and doctors by providing automatically annotated daily health reports. In addition, a huge amount of multimedia data such as videos await to be analyzed and indexed for search purposes, while scientific data such as recordings of animal behavior and evolving brain signals are being collected in the hope to deliver a new scientific discovery about life.

In this dissertation, I present a set of extensions of switching linear dynamic systems (SLDSs) which provide the ability to capture the higher-order temporal structures within data and to produce more accurate results for the tasks such as labeling and estimation of global variations within data. The presented models are formulated within a dynamic Bayesian network formulation along with the inference and learning methods thereof.

The previous state-of-the-art standard SLDSs model the nature of continuous multivariate temporal data under the assumption that the characteristics of complex non-linear temporal sequences can be captured by Markov switching between a set of simpler primitives which are linear dynamic systems (LDSs). Accordingly, the SLDS model provides us with the ability to learn the temporal models from training data and to label novel sequences according to regimes that exhibit different dynamics.

However, the standard SLDS model is lacking in several aspects, which leads to its shortcomings in the scope of the data and the tasks it can handle, and in the quality of the labeling results. First, in terms of the quality of the continuous labeling tasks without known segment boundaries, it produces inaccurate and often over-segmented labeling results because it blindly adopts the geometric duration models implied by the Markov assumption. Second, the standard SLDS model does not provide principled mechanisms to capture or to

infer the amount of global variations within data, which we refer to as the quantification task. Accordingly, they tend to produce both inaccurate labeling and quantification results. Third, it can not effectively model the data with grammar-like hierarchical temporal structure. Accordingly, the standard SLDSs do not provide means to interpret data at multiple temporal or semantic granularities and often produce less than impressive labeling results.

In this dissertation, we address all of the above limitations of standard SLDSs by enhancing the model to incorporate higher-order temporal structures.

First, segmental SLDSs (S-SLDSs) produce superior labeling results by capturing the descriptive duration patterns within each LDS segment. The encoded duration models describe data more descriptively and allow us to avoid the severe problem of over-segmented labels, which leads to superior accuracy.

Second, parametric SLDSs (P-SLDSs) allows us to encode the temporal data with global variations. In particular, we have identified two types of global systematic variations : temporal and spatial variations. The P-SLDS model assumes that there is an underlying canonical model which is globally transformed in time and space by the two associated global parameters respectively. Accordingly, P-SLDSs can solve the quantification problem of estimating the global variations within data and simultaneously produce the labeling results with superior accuracy.

Third, we present hierarchical SLDSs (H-SLDSs), a generalization of standard SLDSs with hierarchic Markov chains. H-SLDSs are able to encode temporal data which exhibits hierarchic structure where the underlying low-level temporal patterns repeatedly appear among different higher-level contexts. Accordingly, H-SLDSs can be used to analyze temporal data at multiple temporal granularities, and provide the additional ability to learn a more complex H-SLDS model easily by combining underlying models.

The developed SLDS extensions have been applied to two real-world problems. The first problem is to automatically decode the dance messages of honey bee dances where the goal is to correctly segment the dance sequences into different regimes and parse the messages about the location of food sources embedded in the data. We show that a combination of the P-SLDS and S-SLDS models has demonstrated improved labeling accuracy and message

parsing (an instance of a quantification task) results. The second problem is to analyze wearable exercise data where we aim to provide an automatically generated exercise record at multiple temporal and semantic resolutions. It is demonstrated that the H-SLDS model with multiple layers can be learned from data, and can be successfully applied to interpret the exercise data at multiple granularities. It is also shown that the H-SLDS model produces superior labeling results than the standard SLDSs for low-level semantic patterns, due to the use of higher-level temporal structure.

Chapter I

INTRODUCTION

We introduce new temporal models to address the challenges in interpreting multivariate time-series data in this dissertation. These models build upon the previously developed standard switching linear dynamic systems (SLDSs) by adding additional model parameters which are designed to encode the higher-level temporal structure often not captured by the original SLDSs. The resulting models can be learned from data, and can be used to label data w.r.t. the dynamics exhibited by data, and to identify underlying global factors which produce the systematic variations within data. The experimental results demonstrate that the proposed models produce superior accuracy over the standard SLDSs for a variety of tasks, guided by the descriptive temporal structure and the learned function of the data variations w.r.t. the global factors.

My thesis in this dissertation is the following :

> Switching linear dynamic systems with higher-order temporal structure increase the scope of the data and the temporal inference tasks that can be handled, and produce superior labeling results over the standard SLDSs.

Specifically, the thesis can be split into three claims that I will defend with the corresponding extensions of SLDSs in this dissertation :

1. Segmental SLDSs (S-SLDSs) produce superior labeling results by capturing the descriptive duration patterns within each LDS segment.

2. Parametric SLDSs (P-SLDSs) can model data with global variations and provides superior labeling accuracy along with the additional ability to estimate the amount of global transformation exhibited by data.

3. Hierarchical SLDSs (H-SLDSs), a generalization of standard SLDSs with hierarchic Markov chains, are able to encode temporal data which exhibits grammar-like hierarchic structure and provides the ability to label temporal data at multiple temporal granularities along with superior labeling accuracy.

In the following sections, we describe the core problems of interest in temporal sequence analysis, our approach towards the problems, and how it advances the state of the art in this field.

1.1 Automated Temporal Sequence Analysis

Temporal sequences are abundant. Examples of temporal data include motion trajectories, voice, video frames, medical sensor signals (e.g., fMRI or heartbeat signals), wearable sensor data, and economic indices, only to name a few. Temporal data in the most general form is a sequence of multivariate vectors.

In contrast to the abundance of temporal data, the analysis of such data still often relies on manual interpretation by humans. The manual interpretation of the temporal data, which is a time-consuming process, seems challenging in some cases due to the complexity of data. For example, sound technicians study complex sound waves, medical doctors conduct diagnosis based on the signals recorded from medical monitoring systems, and investment bank analysts analyze stock price histories. In other occasions, the tasks seem simpler, but, the data has still been mostly interpreted and labeled by humans, simply due to the lack of automated analysis tools. For example, computer graphics experts in animation industry often search for a particular motion sequence from a database investing substantial amount of time, and field biologists label the tracked motion sequences of animals w.r.t. the corresponding motion regimes by thoroughly examining the tracks frame by frame.

We can observe that the development of automated tools to analyze temporal sequences can contribute to such diverse fields where they would assist the knowledge workers to improve their work productivity through the automation of diverse manual works. Moreover, these new tools can provide them with the ability to explore a large temporal sequence database, which was previously challenging due to the substantial amount of manual work

required. In addition, the advances in temporal sequence analysis can contribute to so-called emerging real-time applications which are targeted to proactively respond to the needs of the users based on the sensor signals, e.g., medical monitoring devices and wearable gesture recognition system among many others. Such systems are designed to improve the quality of people's lives by allowing them to have the necessary assistance at all times and help them to conduct their tasks more easily.

In the following sections, we will describe important problems in automated temporal sequence analysis, what are lacking in the state of the art to address such problems, and how we address the challenges by developing models with higher-order temporal structures.

1.2 Model-based Approach : SLDSs

We adopt switching linear dynamic systems (SLDSs), a probabilistic generative model, to characterize temporal sequences based on training data and to interpret novel temporal sequences. The SLDS model assumes that complex temporal sequences can be described by the Markov switching between a set of simpler primitives which are linear dynamic systems (LDSs), often called a Kalman filter model. Hence, the distinctive feature of SLDSs is the fact that they use LDSs as their primitives, compared to the piecewise constant approximators used by the popular hidden Markov models (HMMs) [70]. Accordingly, the SLDS model provides us with the ability to label novel sequences into regimes that exhibit different dynamics.

Due to its appealing characteristic, the SLDS model has been studied in a variety of problem domains. Representative examples include computer vision [68, 67, 69, 59, 16, 2, 63], computer graphics [48, 71], control systems [86], econometrics [41], speech recognition [66, 73], tracking [11], plan recognition [89], machine learning [46, 34, 61, 60, 37], signal processing [25, 26], statistics [80] and visualization [90], to name a few.

In particular, a set of linear dynamic system primitives provide potential advantages over piecewise constant approximators of HMMs in several aspects. First, LDSs can be more descriptive and concise for certain types of temporal sequences. Every LDS encodes the associated dynamics, not measured values, exhibited by the corresponding data. Hence,

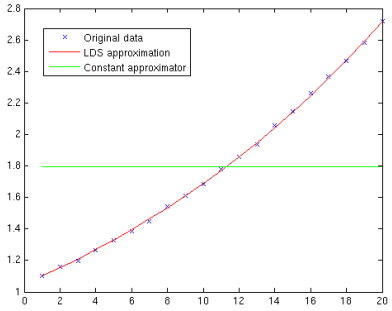


Figure 1: Modeling of a 1D temporal sequence by an LDS and a piecewise approximation. The blue x's indicate the original training sequence changing over time (x-axis). The red and green lines represent the approximation result by an LDS model and a constant approximator. It can be observed that an LDS can model the example dynamic sequence more accurately than a constant approximator.

an LDS has the possibility to describe the primitive patterns of temporal sequences more concisely and accurately in case the data demonstrates certain dynamics. In contrast, the piecewise approximation primitives used by HMMs discard the temporal dynamic information and try to extract key values from data. For example, the 1D sequence in Fig. 1 can be modeled concisely by an LDS while a piecewise approximator will simply learn the mean of the data and fails to capture the distinctive dynamic information.

Additionally, LDS provides the possibility to effectively deal with high-dimensional data using dimensionality reduction techniques. For example, a subset of LDSs are dynamic extensions of the factor analysis model (FA) [12, 35], a dimensionality reduction technique. In certain domains, data may exhibit very high dimensionality, e.g., video data of computer vision community. While a huge number of piecewise approximators will be needed to extract the informative key-points from high-dimensional sequences, relatively small number of LDSs may be needed to model the dynamic patterns from the dimension-reduced data[1].

In terms of temporal ordering structure, the standard SLDSs encode such structure by a Markov transition matrix, identical to HMMs. The Markov assumption simplifies the temporal structure learning problem for switching models by assuming that the short-term

[1]It is worth noting that HMMs are often used with the pseudo-measurements which are obtained through dimension reduction techniques such as PCA from original observations. A more principled generalization such as an HMM extension with underlying factor analysis components [72] has been developed for the speech recognition problems.

switching patterns of discrete modes are descriptive enough to encode the characteristics of data.

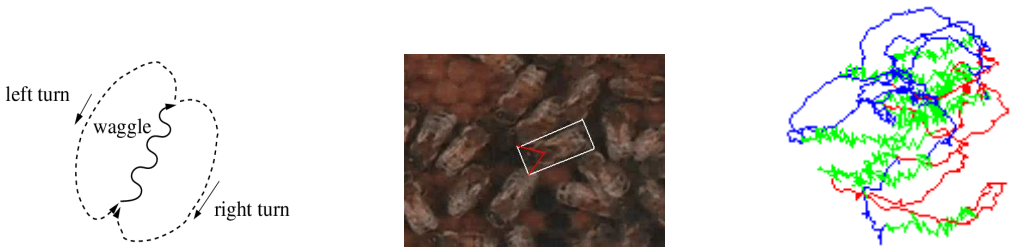
1.3 Beyond standard SLDSs

While the standard SLDSs have the promising properties mentioned in Section 1.2, they are lacking in several aspects, which leads to their shortcomings in the scope of the data and the tasks it can handle, and in the quality of the temporal sequence analysis results. In the following sections, we describe the important tasks in temporal sequence analysis, the limitations of the standard SLDS model to address these tasks, and our novel developments to overcome the challenges.

1.3.1 Superior Labeling of Temporal Data via Duration Modeling

One important task is *label*ing, which is to categorize every part of temporal sequences into different classes based on the properties it exhibits. Labeling of temporal sequences appears in many fields and applications. The classes can be defined by the domain experts based on the semantic concepts they are interested, can be discovered in an unsupervised manner, or may be simply the low-level patterns of dynamics exhibited by data. We can describe the problem more in detail with an application in the biology domain. Honey bees communicate the location and distance to a food source through a stylized dance that takes place within the hive. The dance is decomposed into three different regimes: "left turn", "right turn" and "waggle", as shown in Fig. 2(a). The length (duration) and orientation of the waggle phase correspond to the distance and the orientation to the food source. Figure 2(b) shows a dancer bee that was tracked by a previously developed vision-based tracker [40, 78]. The labeling problem in this domain is to automatically segment the trajectory into one of the three categories. An example result obtained by a developed automated sequence tool is shown in Fig 2(c) with the color-coded motion patterns.

The standard SLDSs are lacking in their duration modeling capability, which often leads to inaccurate labeling results to exhibit over-segmented results. Specifically, the Markov assumption limits SLDSs from capturing descriptive duration patterns of LDSs and only allow the induced geometric duration distributions within each LDS regime. The fact that



(a) (b) (c)

Figure 2: (a) A bee dance consists of three patterns : waggle, left turn, and right turn. (b) The box in the middle is a tracked bee. (c) An example honey bee dance trajectory. The track is automatically obtained using a vision-based tracker and automatically labeled afterwards.
Key : waggle, right-turn, left-turn

LDSs encode dynamics suggests that every LDS is expected to be active for a longer time-span than the piecewise constant approximators of HMMs. However, the geometric duration model induced by Markov assumption assigns the highest probability to the duration of one. Consequently, accurate segmentation results are expected only when the observation is minimally noisy and the discrepancies between the model and the intrinsic characteristics of the data are trivial. In reality, sensory data can possess substantial amount of noise. Accordingly, the simple Markovian assumption can result in labeling results with over-segmentations where false labels with very short durations can be inserted due to the noise at those particular time frames.

In this dissertation, we present *segmental SLDSs* (S-SLDSs), presented in Chapter 3, which is an extension of SLDSs consisting of LDSs with more accurate time duration models. In particular, S-SLDSs effectively discount short-term strong noise and produce more accurate labeling results, guided by the descriptive prior knowledge on the durations of the LDS models.

1.3.2 Modeling and Interpreting Data with Global Variations

The ability to model temporal data with *global* variations and to infer the amount of global variations within novel sequences are very important but relatively less addressed issues in temporal sequence analysis. By global variations, we mean the global factors that underlie the *systematic* variations in the data. For example, the intended pointing direction of a

person changes the overall trajectory of the person's gesture. Another example is the dance trajectories of honey bees which vary substantially depending on the distance and orientation to the indicating food sources. Examples of varying honey bee dance trajectories in stylized forms are shown in Figure 3.

The ability to encode global variations is important not only because it can lead to more accurate labeling results by understanding the context where the data stems from but also because the estimation of the amount of global variations, which we refer to as a *quantification* task, is often the important task of interest. In many cases, we are more interested in the global parameters that vary the behavior of the signals rather than the exact categorization of the sub-regimes. Accordingly, the quantification of global variables provide users with the high-level information they need. In other words, there are data with global parameters which affect the whole sequence and one can model it explicitly and estimate them accordingly. For example, we would be more interested in capturing the pointing direction of a human gesture and the messages about the food locations within the honey bee dances rather than the explicit labeling results.

Most of the temporal models such as standard HMMs and standard SLDSs are lacking to encode the global variations within data because their developments were focused on capturing only the *local* variations within data. For example, it is well known that standard HMMs demonstrate excellent ability to capture the slight variations within speech data. However, once the amount of variations increases, which are often due to the global effects caused by different speakers, standard HMMs need to rely on the use of mixture of observation models or even a set of HMMs in an unprincipled way [70], ignoring the underlying structure of the problems. More importantly, these models do not provide principled mechanisms to estimate the amount of global variations from data. The previous work in HMMs which address the global transformation within data include a parametric HMMs (P-HMM) [87] where the problem of recognizing globally parameterized gestures are studied, and 'Style Machine' by Brand and Hertzmann [14] where they developed a method to vary the styles of dance sequences by treating them to globally condition their HMM model. A related transformation-invariant learning approach for video analysis appeared in [31].

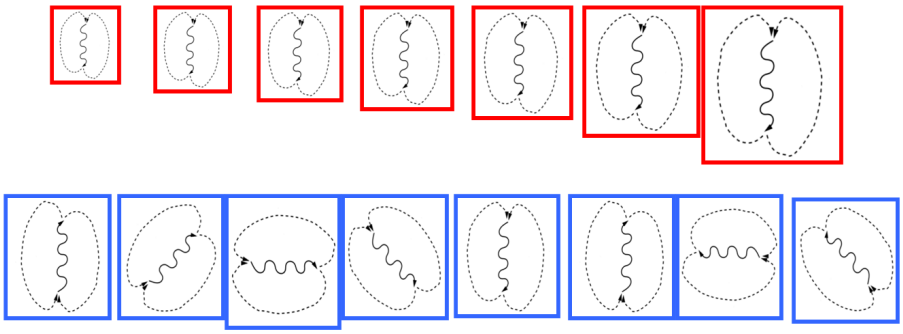


Figure 3: Examples of diverse honey bee dance trajectories in stylized forms. The upper row shows the trajectory variations which depend on the duration of the waggle phase. The bottom row shows the trajectory variations which depend on the orientation of the waggle phase. Waggle phase and angle indicate the distance and orientation to the food source.

In this dissertation, we present an extension of SLDSs, *parametric SLDSs* (P-SLDSs), which assumes that there is an underlying canonical template which is globally transformed based on a set of global transformation factors. In particular, we have identified two types of global systematic variations : temporal and spatial variations. Temporal variations correspond to the rate of dynamic switching within data. On the other hand, the spatial variation refers to the correlated transformation that applies to the overall temporal sequences. Accordingly, the P-SLDS model describes the data with global variations by transforming the canonical template both in space and time. An illustrative example is the honeybee dances : bees communicate the orientation and the distance to the food sources through the dance angles and waggle duration of their stylized dances. In this example, the canonical underlying template is in the form of the prototype dance trajectory illustrated in Fig. 2 (a) where a set of instantiated example trajectories are shown in Fig. 3.

In addition, we introduce an inference method for the P-SLDS model where we solve the labeling problem and the quantification problem simultaneously. The intuition behind the presented inference method is that an accurate estimate on the amount of global transformation will lead to a superior labeling result since it will provide a strong cue for the context of the data. Moreover, it is sensible to expect that the improved low-level labeling results can lead to superior quantification results in return. Specifically, we formulate expectation-maximization (EM) methods for learning and inference in P-SLDS and present it in Chapter

4 of this dissertation. It is demonstrated that P-SLDSs can solve the quantification problem of estimating the global variations within data and simultaneously produce labeling results with superior accuracy.

1.3.3 Modeling and Interpreting Data with Hierarchy

It is very important to be able to model data with a grammar-like hierarchical structure. By data with grammar-like hierarchic structures, we mean the data where the underlying low-level temporal patterns repeatedly appear among different higher-level contexts through certain stochastic temporal ordering processes. In other words, the data has a grammar-like hierarchical structure when (1) the overall behavior of the data over time can be modeled by a probabilistic grammar such as a probabilistic context free grammar where the symbols in the grammar correspond to different sub-states of the system, (2) the symbols in the grammar can be categorized into multiple hierarchies where each hierarchy exhibit distinct levels of abstraction in terms of temporal duration and semantic extent, and (3) the parent-children relationships between the states within adjacent hierarchies can be characterized. As an example, we can see the two different data sequences in Fig. 4 and Fig. 5 where the dynamic patterns of the triangle sides are color-coded. The sequence in Fig. 4 (a) starts with the upward triangle pattern (repeating 10 times) and finishes in the downward triangle pattern (repeating 10 times). In contrast, the training sequence in Fig. 5 (a) starts with the downward triangle pattern and finishes in the upward triangle pattern. On the right side of Fig. 4 and Fig. 5, the transition tables of the corresponding hierarchical Markov models and the flat Markov models are shown. It can be observed that the hierarchical models can capture the long-term temporal structure along with the detailed repetitive patterns, which would allow the two different models to reliably differentiate the two structurally different training sequences. On the other hand, the inability of flat Markov models to capture both the long-term and the repetitive patterns can be shown in the transition tables in Fig. 4 (c) and Fig. 5 (c) where the transition tables are identical, due to the property of Markov models which only learn the averaged switching patterns. It is worth noting that it would be potentially possible to capture such behaviors within data by increasing the number

of states in the flat models, which amounts to the strategy of duplicating an identical state to multiple states w.r.t. the surrounding contexts. However, the true benefits of representational descriptiveness of hierarchical models is that they impose certain transition structure between the states in comparison to the flat models with increased number of states. Accordingly, the hierarchical models can be learned in a modular way and may need fewer training data because the imposed hierarchical structure reduces the structural search space substantially compared to flat models.

Furthermore, the encoding of hierarchic structure within data provides us with the powerful ability to interpret data at multiple semantic and temporal granularities. In particular, most of the semantic labels that we are interested in categorizing are hardly described by a single LDS model. For example, the gym exercise of a person and the different behavioral modes of a multi-robot system exhibit very complex temporal behavior over a substantial span of time which would comprise of large number of primitive patterns. Moreover, such complex behaviors usually comprise of another set of less complex behaviors which builds up on a set of primitive patterns, e.g., LDSs, in a nested manner, forming a hierarchy between the semantic top level concepts down-to low-level primitive dynamic patterns. In certain cases, we would be interested in interpreting the behavior of the signals at all possible semantic and temporal resolutions, a task which can be hardly done in a principled manner when we do not encode such mappings between high-level concepts to low-level signals.

While there has been previous work to develop SLDSs with hierarchic Markov structure [37, 91], they were lacking in some aspect both in theoretical and empirical aspects. From the theoretical point of view, their models do not provide the ability to capture the descriptive call-return semantics. By call-return semantics, we mean that a higher-level concept which initiates a chain of actions at the lower level can only switch to another state when the lower-level states terminate. For example, the triangle sequence in Fig. 4 switches from the upward triangle state to the downward triangle state when only a complete upward triangle is completed. However, the previous hierarchical extensions of SLDSs [37, 91] do not provide such descriptiveness in modeling, and the triangles can switch from one pattern to another even when the first triangle has not been closed. In terms of empirical analysis, the previous

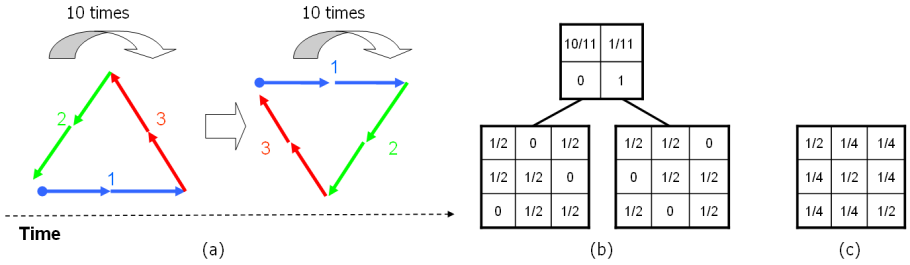


Figure 4: (a) A training sequence. (b) A two-level hierarchical model and (c) a flat Markov model, learned from the training data on the left.

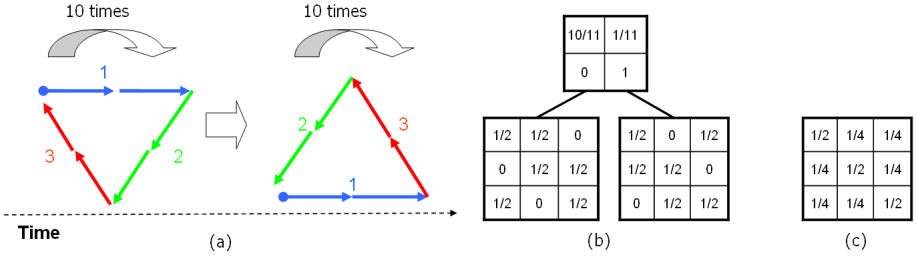


Figure 5: (a) A training sequence. (b) A two-level hierarchical model and (c) a flat Markov model, learned from the training data on the left.

work [37, 91] was applied to either too simple toy problems [91] or they were applied to too complex problems [37] where the detailed analysis of the inference result can be hardly conducted about detailed sub-regimes mostly due to the lack of the ground truth labels.

We present *hierarchical SLDSs* (H-SLDSs) in Chapter 6, a generalization of standard SLDSs with hierarchic Markov chains, as a powerful tool to model and interpret dynamic multivariate data with grammar-like hierarchic structure. H-SLDSs are able to encode temporal data which exhibits hierarchic structure where the underlying low-level temporal patterns repeatedly appear among different higher-level contexts. Accordingly, H-SLDSs can be used to analyze temporal data at multiple temporal granularities, and provide the additional ability to learn a more complex H-SLDS model easily by combining underlying models.

In particular, the developed H-SLDS model provides two main benefits : (1) it allows a set of lower-level behaviors to be easily shared among multiple high-level patterns, effectively increasing scalability and interpretability, and (2) it incorporates additional variables to encode the call-return semantics within data descriptively. As a result, the H-SLDS model

can encode the hierarchic relations between the concepts at different semantic and temporal resolutions, and the correlations between the events appearing far apart, in a succinct and comprehensive manner. On the contrary, the flat Markov models of standard SLDSs can only capture the local temporal correlations. Furthermore, the H-SLDSs allow us to interpret novel data according to the semantic concepts of interest, effectively marginalising out the low-level behaviors of signals modeled based on a set of shared LDSs.

Furthermore, we provide detailed analysis of the empirical results obtained by applying the H-SLDS model to the collected human exercise data. The data are carefully collected from multiple people to analyze the generalization power of the model along with large amount of ground-truth labels at entire hierarchies. Accordingly, the empirical results presented in Chapter 7 provides insight into the H-SLDS model and demonstrates its practical usefulness.

1.3.4 Discussion : Generative and Discriminative modeling

It would be worthwhile to discuss the comparative advantages of using generative models, e.g., SLDSs in our work, over the other class of discriminative modeling approaches [13]. In particular, one of the tasks of primary interest in this dissertation is labeling. It is well known that discriminative models such as conditional random fields (CRFs) [44] are excellent in segmenting and labeling data. For example, the standard CRF model and the variants with hidden temporal structures have been applied to the problems of vision-based gesture recognition [55, 5] where they demonstrated promising results compared to the generative models such as HMMs. In other work, a hierarchical CRF model has been developed to classify the on-going activity and the significant places based on the GPS signals collected from a person [49]. Furthermore, a variant of HMMs which learns discriminative observation models with large margins between class representations has been developed and demonstrated promising results for a speech recognition problem [79]. In particular, the discriminative approaches aim to learn a conditional model $P(X|Z)$ of the targeted hidden variables X given the observed data Z directly. On the other hand, the class of generative models which learn the joint distribution $P(X, Z)$ which is often used in a conditional

posterior form $P(X|Z) \propto P(X)P(Z|X)$ within Bayesian formulation. In the Bayesian formulation of generative modeling, there is always a question about the quality of the imposed form of the prior $P(X)$, which do not occur for the discriminative models. In addition, most generative models assume that the set of temporal observations Z are usually conditionally independent from each other to make the inference problem tractable.

Nonetheless, the generative models provide several advantages over the discriminative models, which leads us to pursue such a direction in this dissertation, as described below.

First, the generative models such as SLDSs easily allow available domain knowledge, e.g., physics, to be incorporated into the models and provide more interpretable forms of parameterization, which may lead to superior inference results when the amount of training data is limited. For example, Ng and Jordan [4] provided an empirical result where they showed that a generative model (naive Bayes classifier) outperforms a discriminative equivalent (logistic regression) when the size of the training data is relatively limited.

Second, the class of generative models provides the possibility for superior scalability over the discriminative models. Specifically, generative models allow different concept classes to be learned independently from each other. On the other hand, the discriminative models should be learned jointly together because they usually pursue to learn the boundaries between the class representations. Accordingly, larger number of concept classes imply further challenges for the discriminative models since the number of boundaries between classes approximately increases quadratically w.r.t. the number of classes while the complexity of generative modeling increases linearly in general.

Third, the class of generative models provides the ability to sort out outlier data. For example, generative models can be used for the tasks such as anomaly detection through threshold-based filtering on the computed likelihoods. On the other hand, discriminatively trained models are likely to perform poorly for such tasks because the training examples at the boundary between classes are generally weighted more heavily during the learning process than the common patterns.

Finally, generative models provide more straightforward ways to combine and to extend the existing models. For example, the P-SLDS model presented in Chapter 4 extends

the standard SLDSs by incorporating global variables into the generative process, and the hierarchical SLDS model presented in Chapter 6 can be formed from an existing set of underlying models through a simple procedure. However, it is not clear how the continuous global variables can be easily fit into the discriminative models such as CRFs. Similarly, the hierarchical CRFs presented in [49, 84] requires substantial amount of additional learning procedure whenever we introduce new hierarchies.

As discussed above, generative models provide several advantages over the class of discriminative models, in particular, for the type of data and the tasks that we are interested in this dissertation. Accordingly, we would regard the comparison of our work presented in this dissertation against the discriminative alternatives as our future work and would not directly compare each other in the forthcoming chapters.

1.4 Summary of Contributions

In summary, we present three theoretical contributions in this dissertation which extend SLDSs to produce superior labeling accuracy over the standard SLDS model for a diverse types of data and tasks :

1. Segmental SLDSs, SLDSs with duration models, and the associated learning and inference algorithms are presented in the Chapter 3. S-SLDSs produce more accurate labeling results than the standard SLDSs as the result of more descriptive duration modeling.

2. Parametric SLDSs, a new representation which explicitly models the deformation function between the canonical template w.r.t. the global parameters, is presented in Chapter 4. P-SLDSs provide a principled way to estimate high-level global parameters from data. The associated learning and inference algorithms are presented in this work where we additionally demonstrate superior labeling accuracy.

3. Hierarchical SLDSs, a hierarchical extension of SLDSs, and the associated learning and inference algorithms are presented in Chapter. 6. H-SLDSs allow us to model data with hierarchic structure and to interpret data at multiple temporal resolutions. Moreover,

it allows us to re-use sub-structures, which leads to the reduction in representational redundancy, and to encode domain knowledge on temporal hierarchy into the model more easily.

The rest of this dissertation describes the above-mentioned extensions in detail with the experimental results that demonstrate their usefulness. We will review the background on the standard SLDS model in Chapter 2 to introduce the background knowledge as well as the notations to be used through this dissertation.

1.5 Declaration of Previous work

This dissertation is based on the following previously published material :

- "Data-Driven MCMC for Learning and Inference in Switching Linear Dynamic Systems" by Sang Min Oh, James M. Rehg, Tucker Balch, Frank Dellaert, Twentieth National Conference on Artificial Intelligence (AAAI-2005), Pittsburgh, U.S.A [61].

- "Learning and Inference in Parametric Switching Linear Dynamic Systems" by Sang Min Oh, James M. Rehg, Tucker Balch, Frank Dellaert, IEEE 2005 International Conference on Computer Vision (ICCV-2005), Beijing, China [62].

- "A Variational inference method for Switching Linear Dynamic Systems" by Sang Min Oh, Ananth Ranganathan, James M. Rehg, Frank Dellaert, Technical Report GIT-GVU-05-16, 2005, Georgia Institute of Technology [60].

- "Segmental Switching Linear Dynamic Systems" by Sang Min Oh, James M. Rehg, Frank Dellaert, Technical Report GIT-CC-05-13, 2005, Georgia Institute of Technology [65].

- "Parameterized duration modeling for switching linear dynamic systems" by Sang Min Oh, James M. Rehg, Frank Dellaert, IEEE 2006 International Conference on Computer Vision and Pattern Recognition (CVPR 2006), NYC, U.S.A [64].

- "Learning and Inferring Motion Patterns using Parametric Segmental Switching Linear Dynamic Systems", by Sang Min Oh, James M. Rehg, Tucker Balch, Frank Dellaert

International Journal of Computer Vision (IJCV), Special Issue on Learning for Vision, May 2008. Vol.77(1-3). Pages 103-124 [63].

Chapter II

BACKGROUND : SWITCHING LINEAR DYNAMIC SYSTEMS

In this chapter, we describe the background knowledge on the standard switching linear dynamic system (SLDS) model which serves as the baseline model in this dissertation. While there are several versions of SLDSs in the literature, this paper addresses the model structure depicted in Figure 7, which we describe in detail in the following sections.

An SLDS model represents the nonlinear dynamic behavior of a complex system by the switching among a set of linear dynamic models over time. In contrast to HMM's [70], the Markov process in an SLDS selects from a set of continuously-evolving linear Gaussian dynamics, rather than a fixed Gaussian mixture density. As a consequence, an SLDS has potentially greater descriptive power. Offsetting this advantage is the fact that exact inference in an SLDS is intractable because the continuous states are coupled during the switchings, which complicates inference and parameter learning [45].

The rest of this chapter is organized as follows. First, we review the linear dynamic systems (LDSs), and its extension, switching LDSs (SLDSs). Then, we review a set of the developed approximate inference techniques and an EM-based learning method for SLDSs. Finally, related work is reviewed and this chapter concludes.

2.1 Linear Dynamic Systems

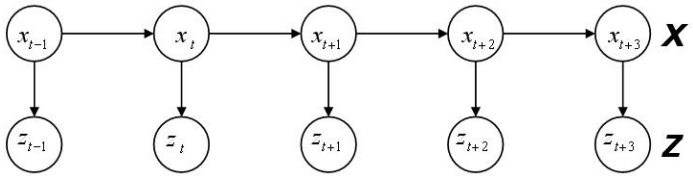


Figure 6: A linear dynamic system (LDS)

A Linear Dynamic System (LDS) is a time-series state-space model consisting of a linear Gaussian dynamics model and a linear Gaussian observation model. The graphical representation of an LDS is shown in Fig. 6. The Markov chain at the top represents the state evolution of the continuous hidden states x_t. The prior density p_1 on the initial state x_1 is assumed to be normal with mean μ_1 and covariance Σ_1, i.e., $x_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$.

The state x_t is obtained by the product of state transition matrix F and the previous state x_{t-1}, corrupted by zero-mean white noise w_t with covariance matrix Q:

$$x_t = Fx_{t-1} + w_t \text{ where } w_t \sim \mathcal{N}(0, Q) \tag{1}$$

In addition, the measurement z_t is generated from the current state x_t through the observation matrix H, and corrupted by zero-mean observation noise v_t:

$$z_t = Hx_t + v_t \text{ where } v_t \sim \mathcal{N}(0, V) \tag{2}$$

Thus, an LDS model M is defined by the tuple $M \stackrel{\Delta}{=} \{(\mu_1, \Sigma_1), (F, Q), (H, V)\}$. Exact inference in an LDS can be performed using the RTS smoother [10], an efficient variant of belief propagation for linear Gaussian models. Further details on LDSs can be found in [10, 50, 74].

Linear dynamic systems have been often used for tracking problems [10, 50]. In addition, LDSs have been used to model the overall texture of video scenes in a compact way with the video as a sequence of observations and generate a infinitely long video similar to the training sequences [24]. In other work, multiple LDSs were used to segment video w.r.t. the associated temporal texture patterns [19].

2.2 Switching Linear Dynamic Systems

In a switching LDS (SLDS) model, we assume the existence of n distinct LDS models $M \stackrel{\Delta}{=} \{M_l | 1 \leq l \leq n\}$. The graphical model corresponding to an SLDS is shown in Fig. 7. The middle chain, representing the hidden state sequence $X \stackrel{\Delta}{=} \{x_t | 1 \leq t \leq T\}$, together with the observations $Z \stackrel{\Delta}{=} \{z_t | 1 \leq t \leq T\}$ at the bottom, is identical to an LDS in Fig. 6. However, we now have an additional discrete Markov chain $L \stackrel{\Delta}{=} \{l_t | 1 \leq t \leq T\}$ that

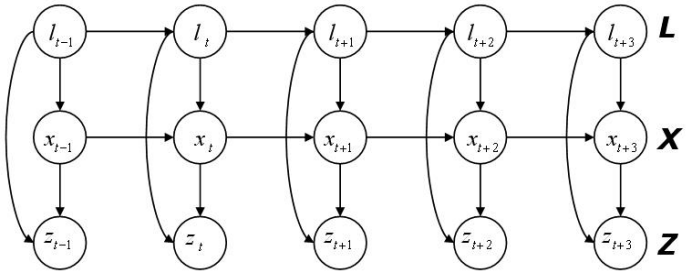


Figure 7: Switching linear dynamic systems (SLDS)

determines which of the n models M_l is used at every time-step. We call $l_t \in M$ the label at time t and L a label sequence.

The switching state l_t is obtained from the previous state label l_{t-1} based on the Markov transition model $P(l_t|l_{t-1}, B)$ which is represented as an $n \times n$ transition matrix B that defines the switching behavior between the n distinct LDS models : $P(l_t|l_{t-1}, B) = B\{i, j\}$.

The state x_t is obtained by the product of the corresponding state transition matrix F_{l_t} and the previous state x_{t-1}, corrupted by zero-mean white noise w_t with covariance matrix Q_{l_t} :

$$x_t = F_{l_t} x_{t-1} + w_t \text{ where } w_t \sim \mathcal{N}(0, Q_{l_t}) \tag{3}$$

In addition, the measurement z_t is generated from the current state x_t through the corresponding observation matrix H_{l_t}, and corrupted by zero-mean observation noise v_t with covariance V_{l_t}:

$$z_t = H_{l_t} x_t + v_t \text{ where } v_t \sim \mathcal{N}(0, V_{l_t}) \tag{4}$$

Finally, in addition to a set of LDS models M, we specify two additional parameters: a multinomial distribution $\pi(l_1)$ over the initial label l_1 : $P(l_1) = \pi\{l_1\}$.

In summary, a standard SLDS model is defined by a tuple $\Theta \triangleq \Big\{\pi, B, M \triangleq \{M_l | 1 \leq l \leq n\}\Big\}$.

It is worth noting that the previous work on SLDSs often adopts simplified versions of the SLDS model described above by introducing different assumptions on parameter tying. Variations include [80, 34, 2] where only the observation models (H and V) are switching, [68, 67, 69, 59] where only the dynamics models (F and Q) are switching with a single observation model, and [66] where all the parameters (F, Q, H, V) are switching but with the additional assumption that the successive continuous states decouple when switching occurs.

Algorithm 1 Learning algorithm for SLDSs based on expectation maximization (EM) method.

1. Initiate a learning process with an initial model parameter tuple Θ^0.
2. **E-step** : Inference to obtain the posterior distribution :

$$f^i(L, X) \triangleq P(L, X|Z, \Theta^i) \tag{5}$$

   over the hidden variables L and X, using the current guess for the SLDS parameters Θ^i.

3. **M-step** : obtain the updated Θ^{i+1} that maximizes the expected log-likelihood :

$$\Theta^{i+1} \leftarrow \underset{\Theta}{\text{argmax}} \ \langle \log P(L, X, Z|\Theta) \rangle_{f^i(L,X)} \tag{6}$$

4. Check convergence via log-likelihood monitoring.
   If converged, stop. Otherwise, go back to Step 2 and repeat.

In the formulation of [66], the originally coupled transition model in Eq. 4 decouples the continuous states between the two successive time-steps whenever the corresponding discrete regimes switches :

$$x_t \sim \mathcal{N}(\mu_{l_t}, \Sigma_{l_t}) \quad \text{when} \quad l_t \neq l_{t-1}$$

In this dissertation work, we adopt the most generic SLDS model without any parameter tying which appeared in [10, 41, 73, 63].

2.3 Inference in SLDS

Inference in an SLDS model involves computing the posterior distribution of the hidden states, which consist of the (discrete) switching states L and the (continuous) dynamic states X. More formally, the inference procedure in SLDSs corresponds to the computation of the posterior $P(L, X|Z, \Theta)$ on the hidden variables which are the label sequence L and the state sequence X, given the observation sequence Z and the known parameters Θ. In application domains such as behavior recognition, the users are typically interested in inferring the switching states L [63], i.e., the labeling problem. On the other hand, the continuous state sequence X is the variable of interest in applications such as tracking or

signal processing. It is worth noting that inference is also the crucial step in parameter learning via the EM algorithm, in addition to its central role in state estimation.

However, it was proven that the exact inference in SLDSs is intractable [45] with the exception of [66] where the authors assumed that the successive continuous states (x_t, x_{t+1}) are decoupled when switching occurs.

Consequently, an array of approximate inference methods have been developed. The approximate inference in SLDSs has focused primarily on three classes of techniques :

1. Stage-wise filtering-based methods such as approximate Viterbi or GPB2 (generalized pseudo-Bayesian estimator of order 2) which maintain a constant representational size for each time step as data is processed sequentially [68, 67, 69, 10].

2. Structured variational methods which approximate the intractable exact model with a tractable, decoupled model [69, 34, 37], including our own work on a variational approximation method for SLDSs with switching observation models [60]. Expectation-propagation [52, 90] belongs to this class of algorithms since it approximates the intractable model by probabilistic densities with tractable constant representational size.

3. Sampling based methods which sample the hidden variables using Monte Carlo techniques [18, 25, 26, 73], including our own work on a Rao-blackwellised data-driven MCMC method [61, 63].

In particular, the stage-wise filtering-based methods produce inference results which have only finite resolution while the other class of methods can represent probabilistic densities upto arbitrary resolution. In this dissertation, we use the following inference methods in our experiments : an approximate Viterbi method [68], a structured variational method [60], and a Rao-Blackwellised data-driven MCMC method[61].

2.4 Learning in SLDS

The maximum-likelihood (ML) parameters $\hat{\Theta}$ for an SLDS model can be obtained using the Expectation Maximization (EM) algorithm [22]. The hidden variables in EM are the

label sequence L and the state sequence X. Given the observation data Z, EM proceeds as described in Algorithm 1.

There, $\langle\cdot\rangle_W$ denotes the expectation of a function or a statistic $(\cdot)$ w.r.t. a distribution W. The E-step in Eq. 5 corresponds to the inference procedure for SLDSs. As mentioned in Section 2.3, it is proven that the exact inference in SLDSs is intractable [45]. Hence, we should rely on one of the approximate inference methods described in Section 2.3 for the E-step, unless we use the variation in [66] where they put strong assumption that the continuous states decouple when switching occurs.

The learning procedure in Algorithm 1 is simplified in the case where the ground truth LDS labels for the training sequences are known. In that case, every LDS parameters are learned separately based on the corresponding parts of the data. Then, the parameters of the discrete process, initial distribution $\pi(l_1)$ and the Markov switching matrix B, are learned separately.

On the other hand, if the ground truth labels are not available, the underlying LDS models should be learned in an unsupervised way. In that case, the inferred labels in Algorithm 1 are used to partition the data which provide the basis to learn distinct LDS models.

2.5 Related Work

The development of SLDSs is closely related with the work on dimensionality reduction [82, 75, 35] for static (non time-series) data as well. In particular, the SLDS model can be thought to be a dynamic parallel of the work on modeling complex static dataset as a mixture of low-dimensional linear models [82, 35]. This analogy can be made since SLDSs aim to model non-linear and high dimensional data as the mixtures of locally linear dynamic systems which switch over time in the latent space.

Chapter III

SEGMENTAL SWITCHING LINEAR DYNAMIC SYSTEMS

3.1 Need for Improved Duration modeling for Markov models

The duration modeling capability of any first-order Markov model is limited by its own assumption upon the transitions between the discrete switching states, i.e., a geometric distribution. As a consequence of the Markov assumption, the probability of remaining in a given switching state follows a geometric distribution :

$$P(d) \quad = \quad a^{d-1}(1-a) \text{ for } d \geq 1 \tag{7}$$

Above, d denotes the duration of a given switching state and a denotes the probability of a self-transition. One consequence of this model is that a duration of one time-step possesses the largest probability mass. This can be seen in Fig. 8 where the red curve depicts the geometric distribution.

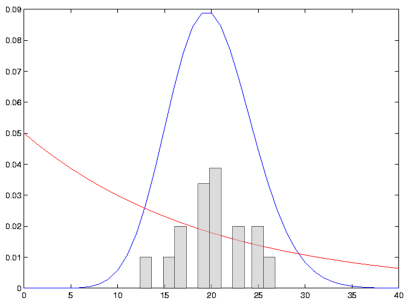


Figure 8: A Gaussian model is closer to the duration distribution of training data (shown as the overlaid histogram) than a geometric duration model.

In contrast, many natural temporal phenomena exhibit patterns of regularity in the duration over which a given model or regime is active. In such cases the geometric duration models of Markov models would not effectively encode the regularity of the data. A honey bee dance is a good example: a dancer bee will attempt to stay in the waggle regime for a certain duration to effectively communicate a message. Another example would be a walking

human. Humans exhibit walking cycles of certain duration which would be modeled better using a Gaussian density. In such cases, it is clear that the actual duration diverges from a geometric distribution.

To illustrate this point, we learned a duration model for the waggle phase using a Gaussian density and a conventional geometric distribution, using one of the manually labeled dance sequences available. Figure 8 shows the learned geometric (red) and Gaussian (blue) distributions for comparison. It can be observed that the Gaussian model is much closer to the training data than the conventional geometric model.

The limitation of a geometric distribution has been previously addressed by the HMM community, and HMM models with enhanced duration capabilities have been developed, which are generally referred to as semi-Markov models [28, 47, 76, 66]. The variable duration HMM (VD-HMM) was introduced in [28] : state durations are modeled explicitly in a variety of PDF forms. Later, a different parameterization of the state durations was introduced where the state transition probabilities are modeled as functions of time, which are referred to as non-stationary HMMs (NS-HMM) [47], sometimes referred to as inhomogeneous or non-homogeneous HMMs. It has since been shown that the VD-HMM and the NS-HMM are duals [23]. In addition, segmental HMM with random effects was developed in the data mining community [33, 42]. Ostendorf et.al. provide an excellent discussion of segmental HMMs in [66].

We adopt similar ideas to arrive at SLDS models with enhanced duration modeling.

3.2 Segmental SLDS

We introduce the segmental SLDS (S-SLDS) model, which improves upon the standard SLDS model by relaxing the Markov assumption at a time-step level to a coarser *segment level*. The development of the S-SLDS model is motivated by the regularity in durations that is often exhibited in nature. For example, as discussed in Section 3.1, a dancer bee will attempt to stay in the waggle regime for a certain duration to effectively communicate the distance to the food source. In such a case, the geometric distribution induced in a standard SLDS is not an appropriate choice. Fig. 8 shows that a geometric distribution assigns the

highest probability to the duration of a single time step. As a result, the label inference in standard SLDSs is susceptible to over-segmentation.

In an S-SLDS, the durations are first modeled explicitly [28] and then non-stationary duration functions [47] are derived from them to be suitably incorporated into the graphical model framework. Both of them are learned from data. As a consequence, the S-SLDS model has more descriptive power and can yield more accurate results than the standard SLDSs. Nonetheless, we show that one can always convert a learned S-SLDS model into an equivalent standard SLDS, operating in a different label space. The approach has the significant advantage of allowing us to reuse the large array of approximate inference and learning techniques developed for SLDSs.

3.2.1 Conceptual View on the Generative Process of S-SLDS

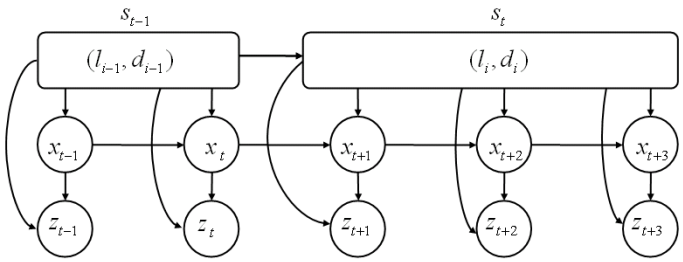


Figure 9: A schematic view of an S-SLDS with explicit duration models.

In an S-SLDS, we deal with segments of finite duration with pre-selected maximum duration D_l^{max} for every label l. Each segment $s_i \triangleq (l_i, d_i)$ is described by a tuple consisting of a label l_i and a duration d_i. Within each segment, a fixed LDS model M_l is used to generate the continuous state sequence for the duration d_i which follows the associated *duration model* D_{l_i}. Similar to SLDSs, we take an S-SLDS to have an $n \times n$ semi Markov label transition matrix $\tilde{B}$ that defines the switching behavior between the segment labels, and an initial distribution $P(l_1)$ over the initial label l_1 of the first segment s_1. The tilde denotes that the matrix is a *semi-Markov* transition matrix between segments rather than between time-steps. Additionally, we associate each label l with a fixed duration model D_l. We denote the set of n duration models as $D \triangleq \{D_l(d) | 1 \leq l \leq n\}$, and refer to them in

what follows as *explicit duration models :*

$$l_i \sim P(l_i|l_{i-1}, \tilde{B}) \quad \text{and} \quad d_i \sim D_{l_i}$$

In summary, an S-SLDS is defined by a tuple $\Theta \stackrel{\Delta}{=} \left\{\pi, \tilde{B}, D, M \stackrel{\Delta}{=} \{M_l|1 \leq l \leq n\}\right\}$.

A schematic depiction of an S-SLDS model is illustrated in Fig. 9. The top chain in the figure is a series of segments where each segment is depicted as a rounded box. In the model, the current segment $s_i \stackrel{\Delta}{=} (l_i, d_i)$ generates a next segment s_{i+1} in the following manner : (1) the current label l_i generates the next label l_{i+1} based on the label transition matrix $\tilde{B}$, (2) then, the next duration d_{i+1} is generated from the duration model for the label l_{i+1}, i.e. $d_{i+1} \sim D_{l_{i+1}}(d)$, (3) the dynamics for the continuous hidden states and observations are identical to a standard SLDS : a segment s_i evolves the set of continuous hidden states X with a corresponding LDS model M_{l_i} for the duration d_i, (4) then the observations Z are generated given the labels L and the set of continuous states X.

3.2.2 Graphical Representation of S-SLDS

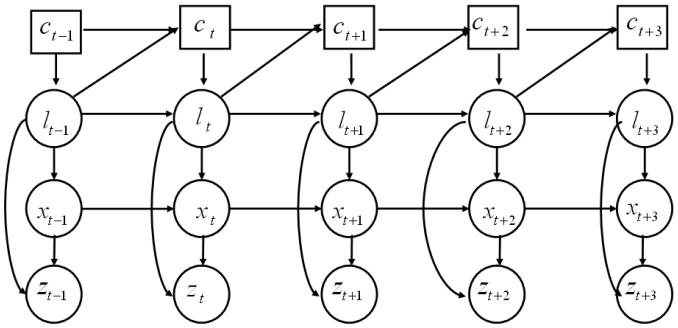


Figure 10: Graphical representation of an S-SLDS

In this section we present a graphical representation of an S-SLDS which transforms the conceptual generative model described in Section 3.2.1 into an equivalent model that uses a conventional Markov switching process at every time-step. To maintain the same duration semantics, we introduce *counter variables* $C \stackrel{\Delta}{=} \{c_t|1 \leq t \leq T\}$. The resulting graphical model of S-SLDS is illustrated in Fig. 10, and is identical to the graphical model of an SLDS, but with additional top chain representing a series of counter variables C.

The variables C maintain an incremental counter which evolves on the basis of *non-stationary transition functions* (NSTFs) $U \stackrel{\Delta}{=} \{U_l(c) | 1 \leq l \leq n\}$. An NSTF U_l for the current label l_t defines the conditional dependency of the next counter variable c_{t+1} given the current counter variable c_t and the label l_t :

$$U_l(c_t) \quad = \quad P(c_{t+1}|c_t, l)$$

The system can either increment the counter, i.e. $c_{t+1} \leftarrow c_t + 1$, or reset it to one, i.e. $c_{t+1} \leftarrow 1$. If the counter variable c_{t+1} is reset, then a label transition occurs, i.e. a new segment is initialized. A new label l_{t+1} is chosen based on the label transition matrix $\tilde{B}$. If the counter simply increments, then the new label is set to be the current label l_t, i.e. $l_{t+1} \leftarrow l_t$.

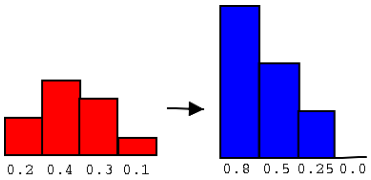


Figure 11: Conversion from explicit duration model D (left) to an equivalent NSTF U (right). As an example, $U(2) = D(2)/\{D(2) + D(3) + D(4)\} = 0.4/0.8 = 0.5$.

We first describe how to convert explicit duration models to equivalent NSTFs. Then, we discuss how the computed NSTFs are used for inference in SLDSs in Sec. 3.4. Given a time series data set, it is straightforward to estimate the parameters of explicit duration models D, as discussed in Sec. 3.2.1. However, in order to incorporate these durations into the SLDS framework, it is necessary to transform the explicit duration models D into equivalent NSTFs U. To do this, we can observe that the explicit duration models D and the NSTFs U are analogous to the duration models of VD-HMMs [28] and NS-HMMs [47] respectively. Hence, we can exploit the duality between VD-HMMs and NS-HMMs, which is described in [23].

The equivalent NSTFs U are evaluated from the explicit duration models D as follows :

$$U_l(c_t) \quad = \quad 1 - \underbrace{\left(D_l(c_t) / \sum_{d=c_t}^{D_l^{max}} D_l(d) \right)}_{\bar{U}_l(c_t)} \tag{8}$$

Above, D_l^{max} denotes the maximum duration allowed for the l th model. Intuitively, the second term $\bar{U}_l(c_t)$ on the r.h.s. in Eq. 8 denotes the probability for a segment with a label l to reset the counter variable, i.e., $c_{t+1} \leftarrow 1$. It represents the ratio of the probability of current duration c_t over the sum of durations equal or greater than c_t in the corresponding duration model D_l. An example is illustrated in Fig. 11 to show the evaluation of an NSTF from an explicit duration model. In summary, an S-SLDS model is completely defined by a tuple $\Theta \triangleq \left\{\pi, \tilde{B}, U \triangleq \{U_l | 1 \leq l \leq n\}, M \triangleq \{M_l | 1 \leq l \leq n\}\right\}$ where the NSTFs U are obtained from the explicit duration models D.

3.3 Learning in Segmental SLDS

Learning in S-SLDSs is analogous to learning in SLDS, using EM. The initial distribution π, and LDS model parameters M are learned in exactly the same manner as in SLDS. However, it is necessary to learn the additional duration models D and the semi-Markov transition matrix $\tilde{B}$. These two additional model parameters can be estimated from the segmental representations of the label sequences L's, i.e., $L = \cup_{j=1}^{|s|} s_j$. The specific functional forms of ML estimation depend upon the choice of duration models. An example duration model would be the Gaussian distribution. However, Gaussian models encode probabilities for non-existing negative durations as well. Hence, only the positive part of the learned Gaussian models were used in our work after normalization. Note that the choice on the form of probability distributions depend on the duration characteristics of data. For example, Gamma or log-normal distributions which only encode probability regions on positive durations can be used.

3.4 Inference for Segmental SLDSs

We describe a convenient inference procedure for S-SLDS which reuses the existing SLDS inference modules by re-parameterizing S-SLDSs into equivalent SLDSs. This is an important advantage as it allows us to readily *reuse* the large array of existing approximate inference algorithms for SLDSs. In other words, the inference in S-SLDSs is identical to that of the standard SLDSs, simply with additional conversion from an S-SLDS to its corresponding SLDS. Note that the conversion algorithm described in this section is an independent

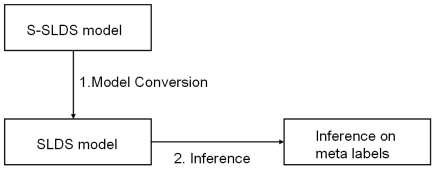


Figure 12: Inference in S-SLDS.

convenience procedure which differs from the conversion for NSTFs described in Eq. 8.

The overall idea of inference is depicted in Figure 12. In step 1, we convert an S-SLDS model into an equivalent SLDS model. Then, we perform step 2 (inference) using any of the approximate inference algorithms for the standard SLDSs. Once the inference results are obtained via available standard SLDS inference methods, the inference results are marginalized to be represented in S-SLDS format and concludes.

The model conversion from an S-SLDSs to an equivalent SLDS is possible by applying the standard technique of merging multiple discrete variables into meta variables. Specifically, all possible pairs of a label l_t and a counter value c_t are merged and form a set of "lc" variables where $\mathcal{LC} \triangleq \{(l, c_i) | 1 \leq l \leq n, 1 \leq c_i \leq D_l^{max}\}$. To obtain a complete SLDS model, an equivalent $n' \times n'$ transition matrix B', where $n' \triangleq \sum_{l=1}^{n} D_l^{max}$, is constructed from the semi-Markov transition matrix $\tilde{B}$ and the NSTFs U, as follows :

$$B'_{(l_i,c_i),(l_j,c_j)} = \begin{cases} U_{l_i}(c_i) & \text{increment} \\ \tilde{B}_{l_i,l_j}(1 - U_{l_i}(c_i)) & \text{reset} \\ 0 & \text{not allowed} \end{cases} \tag{9}$$

In Eq. 9, the three cases for the counter variable differ as follows : (increment) $l_i = l_j$ and $c_j = c_i + 1$, (reset) $c_j = 1$, and (not allowed) for all other cases. In addition, the initial label distribution π' for the equivalent SLDS can similarly be constructed from the S-SLDS initial distribution π :

$$\pi'(l_i, c_i) = \begin{cases} \pi(l_i) & \text{if } c_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

3.4.1 Computational Considerations

Now that we have established that an equivalent SLDS can always be constructed from an arbitrary S-SLDS, we need to consider efficient inference methods. If we reuse the original learning and inference algorithms for SLDSs in a naive manner, the cost of inference will be on the order of $O(TD_{max}^2|L|^2)$ for S-SLDSs, while it takes $O(T|L|^2)$ for SLDSs without duration models, where $D_{max} \stackrel{\Delta}{=} \max_{1\leq l\leq n}\{D_l^{max}\}$ denotes the maximum duration among all labels. Thus, there is a considerable computational overhead, by a factor of $O(D_{max}^2)$. Such increased asymptotic running time overhead applies to all the approximate inference algorithms[1] which require pairwise computations between every time-step in general.

Nonetheless, we can still maintain linear efficiency w.r.t. the maximum duration D_{max} by exploiting the sparseness of the constructed SLDS matrix B'. It can be observed from Eq. 9 that the SLDS matrix B' is mostly sparse, i.e. only a few transitions are allowed between the states in $\mathcal{LC}$. In fact, only $(|L|+1)$ transitions allowed for every lc state. The allowable transitions include the resets to $|L|$ labels and one increment transition. Hence, we can achieve an overall performance of $O(TD_{max}|L|^2)$ via exploiting this fact, which results in reduced overhead by a factor of $O(D_{max})$. The number is derived from the fact that there are total $O(D_{max}|L|)$ states at time $t-1$, and the number of transitions allowed for each state to time t reduces to $O(|L|)$ from $O(D_{max}|L|)$. This reduction in complexity allows us to incorporate a duration model with a large D_{max} and maintain computational efficiency. As a consequence, we can adopt the more powerful duration modeling capabilities of an S-SLDS at the cost of a modest complexity increase over the standard SLDS model.

3.5 Discussion and Future Work

We have presented S-SLDSs which aim to provide additional descriptive duration modeling abilities for each switching state. The S-SLDS model is formulated based on the non-stationary duration models [47] which are dual representations of the explicit duration

[1]Examples include approximate Viterbi methods [69] and variational methods [69, 34, 60] which require the computations between all possible state pairs from the previous time-step to the next time-step. An implementation strategy to naturally exploit the existing sparse structure is to build upon sparse linear algebra libraries.

models [28].

An interesting avenue for future research would be to investigate the use of more compact representations for the duration models to further improve the complexity of the inference algorithms while we maintain the descriptive knowledge about the duration patterns. The current approaches model the durations of sequences given the fixed maximum duration length D_{max} by implementing a left-to-right counter Markov chain without any self-transition or skip (jump), along with the deterministic initialization prior fully on the left-most state. However, this approach increases the computational load by a multiple of D_{max}, which makes the approach less appealing for the domains where the sequences can span over very long durations, e.g., visual surveillance. Such shortcoming typically requires the system designers to sub-sample their data to decrease the maximum duration, which will result in loss of information. Although the conventional approach of full duration modeling used in this dissertation can encode the duration patterns without any loss of information, there may be practical trade-offs between the accuracy in information and resulting computational load (efficiency).

A promising set of solutions seem to be among the class of phase-type distributions [58], e.g., Erlang and Coxian distributions, which aim to approximate the duration data with fewer number of states within a counter Markov chain. For example, Duong et.al. [27] studied the use of discrete Coxian distribution to provide more compact duration models. The discrete Coxian distributions allow self-transitions and allow prior distribution to be non-zero for the states other than the left-most state. In particular, it has been formally proven that the class of Coxian distributions can approximate any duration distributions arbitrarily closely [58]. Erlang distribution does not provide such provable guarantee in the limit, but still provides simpler solutions. In particular, the computational load increases linearly w.r.t. the number of states in Erlang distributions, and quadratically in Coxian distributions. As a result, the following questions still need to be answered about the use of more compact duration model representations : How can we incorporate systematic variables to control such trade-offs between resulting computational loads and the accuracy of information in a principled manner? Fortunately, the phase-type distributions have been studied extensively

within queuing theory and computer system performance modeling communities. It would be interesting to incorporate such work into S-SLDSs in the future.

Chapter IV

PARAMETRIC SWITCHING LINEAR DYNAMIC SYSTEMS

4.1 *Need for the Modeling of Global Variations*

Temporal sequences often exhibit *global variations*, which are often controlled by *global parameters*. For example, people walk, but at different paces and with different styles. Sound fluctuates, but with different frequencies and different amplitudes. Hence, one important problem in temporal sequence analysis is 'quantification', by which we mean the identification of global parameters that underlie the behavior of the signals. However, most switching system models are only designed to be able to label temporal sequences into different regimes, e.g., HMMs or SLDSs do not provide a principled mechanism to conduct quantification. Furthermore, these models are focused on capturing the local variations, not correlated systematic global variations over the entire sequences.

The consideration of global parameters is motivated at least by four observations : (1) temporal sequences can be often described as the combination of a representative template and underlying global variations, (2) we are often more interested in estimating the global parameters rather than the exact categorization of the sub-regimes, (3) accurate quantification results provide possibility to produce superior labeling results by incorporating the context information embedded in the data, and (4) the existing models which do not encode the global factors are suitable to capture the local variations but not the global transformations that produces systematic long-term correlations between measurements, often requiring us to use either a single less accurate model or multiple mixtures of such models.

There has been previous work that tried to incorporate the above-mentioned global transformations into the time-series modeling. For example, Wilson & Bobick addressed this problem by presenting a parametric HMMs (P-HMM) [87]. In P-HMMs, the parametric observation models are conditioned on global observation parameters where a set of globally parameterized gestures such as pointing gestures could be recognized successfully.

P-HMMs have been used to interpret human gestures and demonstrated superior recognition performance in comparison to HMMs.

Inspired by P-HMMs, we extend the standard SLDS model to develop parametric SLDSs (P-SLDSs). As in a P-HMM, the P-SLDS model incorporates global parameters that underlay systematic *spatial* variations of the overall target motion, and is able to estimate the associated global parameters from data. In addition, while P-HMMs only identified global observation parameters which describe the spatial variations in the outputs, we additionally capture global *dynamic* parameters which encode temporal variations. Then, the problems of global parameter quantification and labeling can be solved simultaneously, improving both solutions iteratively. Hence, we formulate expectation-maximization (EM) methods for learning and inference in P-SLDSs, which is presented through this chapter.

4.2 Parametric Switching Linear Dynamic Systems

As discussed in Section 4.1, the standard SLDS model does not provide a principled means to quantify global variations in the motion patterns. For example, honey bees communicate the orientation and distance to food sources through the (spatial) dance angles and (temporal) waggle durations of their stylized dances which take place in the hive, as shown in Fig. 2. As a result, these global motion parameters which encode the messages of the bee dances are the variables of most interest.

In this section, we present a parametric SLDS (P-SLDS) model which makes it possible to (1) model globally parameterized data, (2) quantify the global variables, and (3) solve both labeling and quantification problems simultaneously in an iterative manner based on the EM [22, 51] framework.

4.2.1 Graphical representation of P-SLDS

In P-SLDSs, the discrete state transition probabilities and output probabilities are parameterized by a set of two types of global parameters $\Phi = \{\Phi_d, \Phi_o\}$. The parameters Φ are global in that they systematically affect the entire sequence. The graphical model of P-SLDS is shown in Fig. 13. Note that there are two classes of global parameters : the dynamics parameters Φ_d at the top and the observation parameters Φ_o at the bottom.

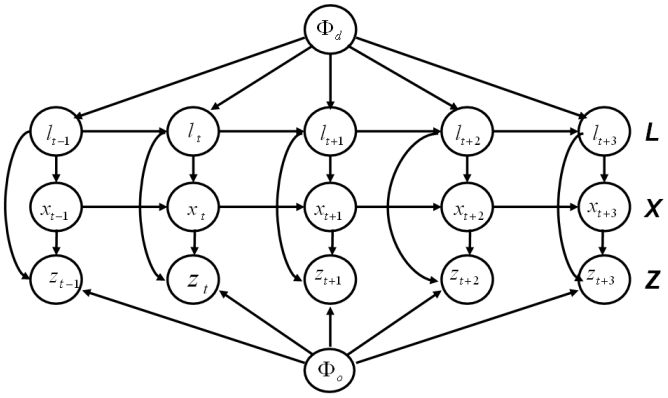


Figure 13: Parametric SLDS (P-SLDS)

The *dynamics parameters* Φ_d represent the factors that cause temporal variations. The different dynamics parameters Φ_d induce distinct switching behavior between behavioral modes. In the case of the honey bee dance, a food source that is far away leads a dancer bee to stay in each dance regime longer, resulting in a dance with larger radius which will show less frequent transitions between the dance regimes. In analogy to the S-SLDS model, the global dynamics parameters are associated with duration models. In contrast, the *observation parameters* Φ_o represent factors that cause spatial variations. A good example is a pointing gesture where the indicating direction changes the overall arm motions. In the honey bee dance case, one can consider a standard SLDS model as a behavioral template that can be stretched in time by the global dynamic parameters Φ_d and spatially rotated by the global observation parameters Φ_o.

The common underlying behavioral template is defined by the canonical parameters Θ. Note that the canonical parameters Θ are embedded in the conditional dependency arcs in Fig. 13. In the bee dancing example, the canonical parameters describe the prototyped stylized bee dance. Then, the individual dynamics in the different bee dances systematically vary from the prototyped dance due to the changing food source locations which are represented by the global parameters Φ.

Notice that the discrete state transitions in the top chain of Fig. 13 are instantiated by Θ and Φ_d, and the observation model at the bottom is instantiated by Θ and Φ_o while the continuous state transitions in the middle chain are instantiated solely by the canonical

parameters Θ. In other words, the dynamics parameters Φ_d, vary the prototyped switching behaviors, and the observation parameters Φ_o vary the prototyped observation model. The intuition behind the quantification of global parameters is that they can be effectively discovered by finding the global parameters that best describe the discrepancies between the new observations and the behavioral template. In other words, the global parameters are estimated by minimizing the residual error that remains between the template and the observation sequence.

The result of *parameterizing* the SLDS model is the incorporation of additional conditioning variables in the initial state distribution $P(l_1|\Theta, \Phi_d)$, the discrete state transition table $P(l_t|l_{t-1}, \Theta, \Phi_d)$, and the observation model $P(z_t|l_t, x_t, \Theta, \Phi_o)$. There are three possibilities for the nature of the parameterization: (a) the PDF is a linear function of the global parameters Φ, (b) the PDF is a non-linear function of Φ, and (c) no functional form for the PDF is available. In the latter case of (c), general function approximators such as a neural network may be learned from data, as suggested in [87]. In Sections 4.3 and 4.4, we discuss learning and inference in P-SLDS under the assumption that functional forms are available. Additionally, we assume that the global parameters are available as part of the training data during the learning phase. However, during the testing phase, we are given only the observation sequence and we estimate the global parameters Φ jointly with the label sequence L.

4.3 Learning in P-SLDS

In the learning phase, P-SLDS learns a canonical behavior template from the data sequences where the individual dynamics may vary due to the different underlying global parameters, but we assume that these parameters are provided as part of our training data. Learning in P-SLDSs entails the estimation of the P-SLDS canonical parameters Θ, given the data $\bar{D} \triangleq \left\{\bar{\Phi} = \{\bar{\Phi}_d, \bar{\Phi}_o\}, \bar{L}, \bar{Z}\right\}$ where the training data $\bar{D}$ comprises a set of known global parameters $\bar{\Phi} = \{\bar{\Phi}_d, \bar{\Phi}_o\}$, ground truth label sequence $\bar{L}$, and the observation sequence $\bar{Z}$. The upper bars $(\bar{\cdot})$ on the variables are used to indicate that the values are known for clarification purposes. We employ EM [22, 51] with the continuous state X as the only hidden

Algorithm 2 EM1 for Learning in P-SLDS

- E-step 1: obtain the posterior distribution

$$f_L^i(X) \triangleq P(X|\Theta^i, \bar{D}) \tag{10}$$

  over the hidden state sequence X, based on the current estimate of the canonical parameters Θ^i.

- M-step 1: maximize the expected log-likelihood :

$$\Theta^{i+1} \leftarrow \underset{\Theta}{\text{argmax}} \ \left\langle \log P(\bar{L}, X, \bar{Z}|\Theta, \bar{\Phi}) \right\rangle_{f_L^i(X)} \tag{11}$$

variables to be inferred during the ML estimation process for the canonical parameters $\hat{\Theta}$. The overall EM algorithm for learning in P-SLDSs is shown in Algorithm 2.

The E-step in Eq. 10 is equivalent to inference in an LDS model. In more detail, since the global parameters $\bar{\Phi}$, the current P-SLDS parameters Θ^i, the label sequence $\bar{L}$, and the observations $\bar{Z}$ are all known, inference over the continuous hidden states X in E-step can be conducted through Kalman-smoothing [9] exactly. Given the posterior distribution $f_L^i(X)$ estimated during E-step (Eq. 10), we then update (re-estimate) the parameters Θ^{i+1} through M-step (Eq. 11).

In the case where the parameterized dependencies such as the Markov switching model $P(l_t|l_{t-1}, \Theta, \Phi_d)$ are linear functions of the global parameters Φ, the M-step in Eq. 11 can be solved analytically. However, in the case where the parametric dependencies are non-linear, an exact M-step is infeasible and needs be solved by alternative optimization methods, e.g., conjugate gradient or Levenberg-Marquardt methods.

4.4 Inference in P-SLDS

We use the learned P-SLDS canonical parameters Θ to quantify the global parameters Φ and infer the label sequence L , given the observations $\bar{Z}$. Note that the canonical parameter set Θ is left to be fixed once they are learned from the training dataset, and we now interpret a novel dataset via inference where neither the global parameters Φ nor the label sequence L are known (hence, upper bars are omitted).

We use EM to quantify the optimal global parameters Φ as shown in Algorithm 3. Note

Algorithm 3 EM2 for Inference in P-SLDS

- E-step 2 : obtain the posterior distribution:

$$f_I^i(L, X) \triangleq P(L, X|\bar{Z}, \Theta, \Phi^i) \tag{12}$$

  over the hidden label sequence L and the state sequence X, using a current guess for the global parameters Φ^i.

- M-step 2 : maximize the expected log-likelihood:

$$\Phi^{i+1} \leftarrow \underset{\Phi}{\text{argmax}} \ \left\langle \log P(L, X, \bar{Z}|\Theta, \Phi) \right\rangle_{f_I^i(L,X)} \tag{13}$$

that we use Algorithm 2 to learn the canonical model parameters Θ, while Algorithm 3 is used to estimate the global parameters Φ with simultaneous inference of the labels L. The details on the EM algorithm in Algorithm 3 are described below. Finally, we will describe a set of helpful strategies to conduct the general EM-based inference, along with our findings about the necessary conditions for successful quantification of the global parameters and their behaviors during the EM-based inference procedure.

4.4.1 E-step 2

The exact E-step in Eq. 12 is proved to be intractable [45], which requires us to rely on the approximate inference methods. Here, we present a derivation of E-step based on approximate Viterbi (VI) method [67]. Note that our derivation can be extended straightforwardly to the other approximate inference methods as well. At the i-th EM iteration, the approximate Viterbi methods approximates the joint posterior over the hidden variables L and X by a single Viterbi label sequence $\hat{L}^i$ and a series of peaked Gaussian distributions over X :

$$\begin{aligned} P(L, X|\bar{Z}, \Phi^i) &= P(X|L, \bar{Z}, \Phi^i) P(L|\bar{Z}, \Phi^i) \\ &\approx P(X|\hat{L}^i, \bar{Z}, \Phi^i) \delta(\hat{L}^i) \end{aligned} \tag{14}$$

Accordingly, the posterior in Eq. 12 during E-step can be re-written as follows :

$$f_I^i(X) \triangleq P(X|\hat{L}^i, \bar{Z}, \Phi^i) \delta(\hat{L}^i)$$

Above, the implicit conditional dependence on the fixed canonical parameters Θ is omitted

for brevity.

4.4.2 M-step 2

Using the approximate posterior $f_I^i(X)$ obtained in Eq. 14, the expected complete log-likelihood $\mathbb{L}^i(\Phi)$ in Eq. 13 is approximated as:

$$\begin{aligned} \mathbb{L}^i(\Phi) &\triangleq \sum_L \int_X \log P(L, X, \bar{Z}|\Phi) P(L, X|\bar{Z}, \Phi^i) \\ &\approx \int_X \log P(\hat{L}^i, X, \bar{Z}|\Phi) f_I^i(X) \end{aligned} \tag{15}$$

Using the chain rule, the first factor in the r.h.s. of Eq. 15 can be factored as :

$$P(\hat{L}^i, X, \bar{Z}|\Phi) = P(\hat{L}^i|\Phi_d) P(X, \bar{Z}|\hat{L}^i, \Phi_o) \tag{16}$$

Note that we now only condition on the relevant global parameters, e.g. the label sequence $\hat{L}^i$ is only conditioned on Φ_d. Substituting Eq. 16 into the expected log-likelihood $\mathbb{L}^i(\Phi)$ in Eq. 15, we obtain a more succinct form of $\mathbb{L}^i(\Phi)$ in which the term $\log P(\hat{L}^i|\Phi_d)$ is moved outside the integral :

$$\begin{aligned} \mathbb{L}^i(\Phi) &= \log P(\hat{L}^i|\Phi_d) + \int_X \log P(X, \bar{Z}|\hat{L}^i, \Phi_o) f_I^i(X) \\ &= \mathbb{L}^i(\Phi_d) + \mathbb{L}^i(\Phi_o) \end{aligned} \tag{17}$$

Above, we introduced two convenience terms, the dynamic log-likelihood $\mathbb{L}(\Phi_d)$ and the observation log-likelihood $\mathbb{L}(\Phi_o)$:

$$\mathbb{L}^i(\Phi_d) \triangleq \log P(\hat{L}^i|\Phi_d) \tag{18}$$

$$\mathbb{L}^i(\Phi_o) \triangleq \int_X \log P(X, \bar{Z}|\hat{L}^i, \Phi_o) f_I^i(X) \tag{19}$$

The factorization in Eq. 17 provides us with a crucial insight for the solution for the learning problem in P-SLDSs : the total expected log-likelihood $\mathbb{L}^i(\Phi)$ is maximized by *independently* updating the global observation parameters Φ_o and dynamic parameters Φ_d, i.e. we obtain the updated global parameters Φ_d^{i+1} and Φ_o^{i+1} by maximizing the dynamic log-likelihood $\mathbb{L}^i(\Phi_d)$ and the observation log-likelihood $\mathbb{L}^i(\Phi_o)$ respectively.

We can further factorize the dynamic log-likelihood $\mathbb{L}^i(\Phi_d)$ in Eq. 18 and the observation log-likelihood $\mathbb{L}^i(\Phi_o)$ in Eq. 19. Finally, we obtain the fully factorized log-likelihood terms :

$$\mathbb{L}^i(\Phi_d) = \log P(\hat{l_1}^i|\Phi_d) + \log \sum_{t=2}^{|Z|} P(\hat{l_t}^i|\hat{l}^i_{t-1}\Phi_d) \quad (20)$$

$$\begin{aligned}\mathbb{L}^i(\Phi_o) &= \int_X \log \left\{ P(\bar{Z}|X, \hat{L}^i, \Phi_o) P(X|\hat{L}^i) \right\} f^i_I(X) \\ &\equiv \int_X \log P(\bar{Z}|X, \hat{L}^i, \Phi_o) f^i_I(X) \\ &= \sum_{t=1}^{|Z|} \int_{x_t} \log P(\bar{z}_t|x_t, \hat{l_t}^i, \Phi_o) f^i_I(x_t) \end{aligned} \quad (21)$$

Above, the term $f^i_I(x_t)$ denotes the marginal distribution on x_t from the full posterior, i.e., $f^i_I(x_t) \stackrel{\Delta}{=} \int_{X/x_t} f^i_I(X)$. Note that the term $\int_X \log P(X|\hat{L}^i) f^i_I(X)$ disappears in the second line of Eq. 21 as it is not a function conditioned by the global observation parameter Φ_o and does not help to improve the observation likelihood $\mathbb{L}^i(\Phi_o)$. In the case where we are modeling data with parametric S-SLDS models (see Chapter 3), the global dynamic parameters Φ_d are associated with the duration models of S-SLDSs, and Eq. 20 is not directly applicable because label transitions occur between segments. Hence, once we obtain the Viterbi labels $\hat{L}^i$, the label sequence is converted into a list of segments, i.e., $\hat{L}^i = \cup_{j=1}^{|s|} s_j$ where $s_j \stackrel{\Delta}{=} (l_j, d_j)$, as described in the chapter for S-SLDSs. Then, the dynamic log-likelihood for parametric S-SLDSs can be evaluated as follows :

$$\begin{aligned}\mathbb{L}^i(\Phi_d) &= \sum_{j=1}^{|s|} \log P(s_j|\Phi_d) \\ &\equiv \sum_{j=1}^{|s|} \log D_{l_j}(d_j) \end{aligned} \quad (22)$$

The observation log-likelihood for parametric S-SLDSs would be evaluated as in Eq. 21. Note that Eq. 22 is derived under the assumption that only the duration models in parametric S-SLDSs are parameterized, not the Markov switching patterns between the segments. The details of the M-step will depend upon the application domain and the functional forms. In the case where the parametric forms are linear in the global parameters Φ, the M-step is obtained analytically. Otherwise, alternative optimization methods can be used to maximize the non-linear log-likelihood function in Eq. 21, as described in Section 4.3.

4.5 Discussion

4.5.1 Initial conditions for Global parameters

Through the work in this dissertation, it has been found that fairly good initial values are needed for the global observation parameters to reach an accurate outcome. In particular, it was the case when the functional dependencies between the global observational parameters and the canonical models are non-linear (e.g., rotation), leaving the optimization landscapes complex. Accordingly, there may be multiple local minima where the proposed EM algorithm would not pass through in case the initial values are quite distinct from the true values in the signals.

On the other hand, such sensitivity of outcome regarding the initial condition was only rarely observed for the global dynamics parameters. It is conjectured that most of the widely-used duration model densities we tried were linear w.r.t. the global dynamics parameters, resulting in fairly straightforward likelihood surfaces for optimizations.

4.5.2 Strategies for the EM-based inference updates

We have frequently observed that there are certain tendencies that the global dynamics parameters tend to converge to smaller values, resulting in faster switching patterns than ground truth values in many cases. While the exact cause is not clearly identified, it is conjectured that the duration densities prefer to be in the forms of concentrated probability mass with high peaks rather than the wide-spread densities with shallow volumes. In particular, the shrinking behavior appeared more prominently when global observation parameters were relatively inaccurate and yields intermediate labeling results which tend to include incorrect labels, e.g., over-segmentations. In the presence of over-segmentations, global dynamics parameters are inaccurately estimated to be in ranges of smaller values.

The effective strategy we found is to update two types of global parameters asynchronously during the EM iterations where we update global observation parameters more frequently than the global dynamics parameters. In detail, in the spirit of generalized EM [57], we do not necessarily have to conduct full M-step at every iteration for both global parameters. Hence, we update global observation parameters until they converge and only

update global dynamics parameters afterwards. The updated dynamics parameters change the overall parameterization of P-SLDSs and creates new room to improve global observations parameters. In other words, a series of M-steps with only global observation parameter updates were conducted after which an intermittent M-step with the global dynamics parameters was conducted. This approach was found to be useful for our application described in Chapter 5.

4.5.3 Priors for the global parameters

The use of probabilistic priors for the global parameters is an interesting issue for P-SLDSs. Nonetheless, we have not identified clear benefits yet. Noting that our current formulation of P-SLDSs assumes no priors over the global parameters, it seems to be natural to include more powerful prior densities for the global parameters. Nonetheless, the prior factors are conjectured to be useful only to limited cases of global dynamics parameters at this moment for the following reasons.

The usefulness of the priors for the global observation parameters would be very limited in general, since the global parameters are linked to a large number of measurements variables. Hence, the inference on the global observation parameters would be mainly driven by the large number of likelihood product terms formed by the links between the global observation parameter and the measurements at every time-step, substantially repressing the sole prior term to be useless. The critical factor that would drive the accurate estimation of the global observation parameters would be the use of excellent initial guesses, and the usefulness can be increased by initiating iterative inference procedures multiple times starting from different initial guesses.

The use of the prior density for the global dynamics parameters is conjectured to be potentially useful to avoid the shrinking behavior of duration models that can appear during the EM iterations for the inference in P-SLDSs. In particular, we can use priors to bias the duration patterns of switching models away from being over-shrunken, to avoid the extreme preference for the fast switching patterns. In other words, the global dynamics parameters can be designed to include two types of duration parameters such as mean and variance,

and the prior can play as regularizers that tend to constrain the variances from being overly shrunken. However, we did not see the practical needs for the priors when we used the strategies for the inference tasks described in Sec. 4.5.2, and it is clear that the usefulness of the priors will be again limited when the length of the time-series data increases and there will be increased number of label segments. The interesting future research question is to study whether the use of priors for the global dynamics parameters would eliminate or lessen the needs for the asynchronous update strategies for EM-based inference described in Sec. 4.5.2.

Chapter V

AUTOMATED ANALYSIS OF HONEY BEE DANCES USING A PARAMETRIC SEGMENTAL SLDS MODEL

In this chapter, we experimentally evaluate the developed theory of S-SLDSs (Ch. 3) and P-SLDSs (Ch. 4) on the real-world honey bee dance dataset [1] for the labeling and quantification tasks. To take advantage of both models, we combine the models and use the resulting parametric segmental SLDS (PS-SLDS) model to learn the temporal patterns from data and use the learned model to conduct labeling and quantification tasks. The details of the parameterization within the model is described in this chapter. The resulting PS-SLDS model demonstrates superior accuracy over the standard SLDS model for both the labeling and the quantification tasks.

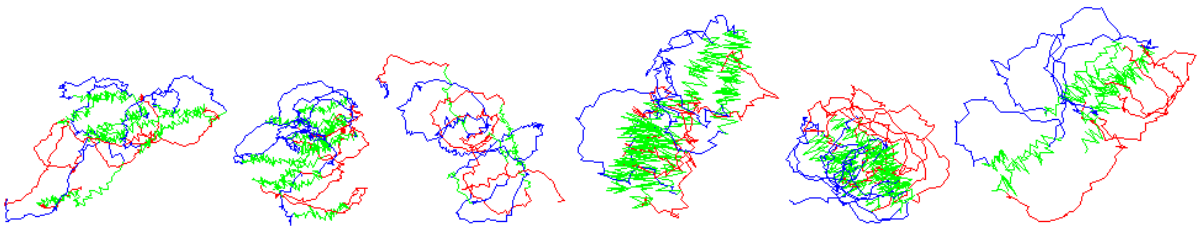
5.1 Motivation

(a) (b) (c)

Figure 14: (a) A honey bee dance consists of three patterns : waggle, left turn, and right turn. (b) A photo of a honey bee hive managed by the researchers at Georgia Tech. (c) A snapshot of a visual tracking system operating on the beehive videos. The white box in the middle is a tracked bee. Examples of honey bee dance trajectories can be seen in Figure 15.

The application domain which motivates the work in this chapter is a new research area which enlists visual tracking and AI modeling techniques in the service of biology [7, 8,

[1] The honey bee dance dataset is publicly available for research purposes. It can be downloaded from : `http://www.cc.gatech.edu/~borg/ijcv_psslds/`.



(1) (2) (3) (4) (5) (6)

Figure 15: Honey bee dance sequences used in the experiments. The trajectories are obtained automatically as the outputs of vision-based trackers. Tables 1 shows the global parameters for each of the numbered sequences.
Key : waggle, right-turn, left-turn

Table 1: The orientation angle (in radian) and duration (in frame) associated with the dataset (sequence numbers refer to Fig. 15). The clockwise angles are measured with zero corresponding to the positive x-axis. The videos were recorded at 30 fps.

Sequence	1	2	3	4	5	6
Orientation angle (radian)	-0.30	-0.25	1.13	-1.33	-2.08	-0.80
Duration (frame)	51.6	46.6	21.4	41.1	19.4	32.6

15, 77, 85]. The current state of biological field work is still dominated by manual data interpretation - a time-consuming process. Automatic interpretation methods can provide field biologists with new tools for the quantitative study of animal behavior.

A classical example of animal behavior and communication is the honey bee dance [32], depicted in a stylized form in Fig. 14(a). Honey bees communicate the location and distance to a food source through a dance that takes place within their hive (an example beehive is shown in Fig. 14(b)). The dance is decomposed into three different regimes: "left turn", "right turn" and "waggle". The length (duration) and orientation of the waggle phase correspond to the distance and the orientation to the food source. Figure 14(c) shows a dancer bee that was tracked by a previously developed vision-based tracker.

The whole six trajectories obtained by the trackers are shown in Figure 15 (1-6) [2]. After tracking, the obtained trajectories of the dancing bees are manually labeled as "left turn" (blue), "right turn" (red) or "waggle" (green), and the associated orientation angles and the durations are estimated, as shown in Table 1. The manually marked labels and the global

[2]For tracking, two different visual tracking systems have been used where the first three sequences (1-3) were tracked by [40] and the latter three sequences (4-6) were obtained by [78].

parameter estimates are to be used as ground truth during the test phases to measure the accuracy of the developed modeling framework.

In this domain, our work on SLDSs is in support of three goals for automatic bee dance analysis. First, we aim to learn the motion patterns of honey bee dances from the labeled training sequences. Second, we should be able to automatically segment novel sequences into the three dance modes reliably, i.e., the labeling problem. Finally, we face a quantification problem where the goal is to automatically deduce the message communicated, in this case : the distance and orientation to the food source. Note that both the labels and the global parameters are unknown, hence the problem is one of simultaneously inferring these hidden variables using the developed PS-SLDS model.

5.1.1 Related Work

There was another effort which aimed to build a system which automatically track honey bees and interpret their behaviors [85] where they used three-layer hierarchical HMMs with an integrated visual tracker that encodes both shapes and appearance of the bees. In their work, they combined ARMA models with the H-HHM model where they track the dancing bees and conduct behavioral analysis simultaneously. The authors reported highly accurate recognition rates on waggle regimes on two sequences where they used higher definition with higher resolution. It is not clear whether leave-one-out approach was used between the two sequences or over-segmentations exist in their results.

In other work, the honey bee data collected by us has been used to empirically validate the usefulness of algorithms developed for the problems such as change detection [88], and models such as SLDSs with hierarchical Dirichlet process prior [30].

5.2 Modeling of Honey bee dances using PS-SLDS

We describe a model for the honey bee dance based on our parametric segmental SLDS (PS-SLDS) model, a combination of the P-SLDS and the S-SLDS models. We expect that the following characteristics of honey bee dance datasets can be captured by PS-SLDSs : (a) the global spatial and temporal variations of the honey bee dances occurring due to the distinct food source locations and (b) the non-exponential characteristics of the duration

patterns of the dance regimes.

The bee dance is parameterized by both classes of global parameters : dynamic parameters for the dance durations and observation parameters for the dance orientations. A set of global dynamics parameter set $\Phi_d \triangleq \{\Phi_{d,l} | 1 \leq l \leq n\}$ is chosen to be correlated with the average duration of each dance regime, where $n = 3$ while a global observation parameter Φ_o is chosen to be the angle orientation of the bee dance.

5.2.1 Canonical parameters

To define the fixed parameters for the canonical templates, we assume the following properties for the honey bee dances :

- The segmental Markov switching parameters between different dance modes are fixed across distinct dances.
- The variance of durations within each behavioral mode is fixed across sequences.
- The dynamics exhibited within each behavioral mode is fixed across distinct bees.

Consequently, the canonical parameters in honey bee dances comprise of a tuple of initial label distribution π, semi-Markov segmental Markov transition matrix $\tilde{B}$, LDS model parameters M and variances in durations in each behavioral modes Σ :

$$\Theta \triangleq \left\{ \pi, \tilde{B}, M \triangleq \{M_l | 1 \leq l \leq n\}, \Sigma \triangleq \{\Sigma_l | 1 \leq l \leq n\} \right\}$$

Note that the canonical parameter tuple Θ is fixed once it is learned from data, as mentioned in Section 4.4 for P-SLDSs. The choice of canonical parameters are based on the knowledge of honey bee dances [32]. For example, it is reasonable to assume that the initial label distribution π and the segment label transition matrix $\tilde{B}$ between different dance regimes do not vary across the dance sequences. In addition, the dancer bees try to regulate their waggle durations to convey the dance messages effectively. Hence, the amount of variation in the duration of each dance regime is assumed to be constant. Hence, they are learned and represented as the variances Σ.

5.2.2 Dynamics model

We set the global dynamic parameters of the l-th model $\Phi_{d,l}$ to be the average duration μ_l of the l-th dance regime, i.e., $\Phi_d \stackrel{\Delta}{=} \{\mu_l | 1 \leq l \leq n\}$. Accordingly, each parameterized duration model D_l of (P)S-SLDSs is modeled as a Gaussian distribution as follows :

$$D_l(c_t) \quad = \quad \mathcal{N}(\mu_l; \Sigma_l) \tag{23}$$

Above, the duration mean μ_l is a global dynamic parameter which is re-estimated at every EM iteration in P(S)-SLDS learning (described in Section 4.4) while the variance Σ_l is a fixed canonical parameter. Then, the explicit duration model in Eq. 23 is used as a discretized histogram with maximum duration length $D_l^{max} = 100$. In the video database, a dance regime with extremely long duration lasted for about 75 frames. Thus, the choice of the maximum duration length D_l^{max} would be sufficient to represent the duration model. Once the histogram duration model D_l is learned, we convert the model into an NSTF U_l , as discussed in Section 3.2.2 on S-SLDSs.

The M-step update for a dynamics parameter $\Phi_{d,l}$ can be obtained by differentiating the dynamic log-likelihood in Eq. 24 :

$$\begin{aligned}
\mathbb{L}^i(\Phi_d) &= \sum_{j=1}^{|s|} \log D_{l_j}(d_j) \\
&= \sum_{l=1}^{N} \left(\sum_{\forall l_j = l} \log D_l(d_j) \right) \\
&= -\frac{1}{2} \sum_{l=1}^{N} \left(\sum_{\forall l_j = l} \log \Sigma_l + \frac{(d_j - \mu_l)^2}{\Sigma_l} \right)
\end{aligned} \tag{24}$$

$$\frac{\partial \log P(\hat{L}|\Phi_d)}{\partial \mu_l} \quad = \frac{2 \sum_{\forall l_j = l} (d_j - \mu_l)}{\Sigma_l} \quad = 0$$

$$\mu_l^{new} \quad \leftarrow \quad \frac{\sum_{\forall l_j = l} d_j}{|s_l|} \tag{25}$$

In fact, the M-step update in Eq. 25 for the global dynamic parameters $\mu^{new} \stackrel{\Delta}{=} \{\mu_l^{new} | 1 \leq l \leq n\}$ turns out to be equivalent to re-estimating the mean durations of distinct dance phases from the estimated segmented label sequence $\hat{L}^i = \cup_{j=1}^{|s|} s_j$.

5.2.3 Observation model

The observation data are time-series sequences of vectors $z_t = [x_t, y_t, \cos(\theta_t), \sin(\theta_t)]^T$where (x_t, y_t) and θ_t denote, respectively, the 2D coordinates and the heading angle of the tracked dancer bee at time t. The angle of zero corresponds to the direction of the positive x-axis and increases in the clockwise direction. The trigonometric function elements in the observations were introduced to bound the effects of angular factors within $[-1, 1]$, to eliminate the boundary condition that appears when the raw radian angles are used. Note that the observed temporary heading angle θ_t differs from the global dance angle Φ_o.

We use the following parameterized observation model $P(z_t|l_t, x_t, \Phi_o)$:

$$z_t \quad \sim \quad \mathcal{N}(R(\Phi_o)H_{\hat{l}_t}x_t, V_{\hat{l}_t}) \tag{26}$$

where $R(\Phi_o)$ is the rotation matrix, and $H_{\hat{l}_t}$ and $V_{\hat{l}_t}$ denote the observation parameters of the $\hat{l}_t$-th component LDS, corresponding to label $\hat{l}_t$ of the Viterbi sequence $\hat{L}$. We also define $\alpha_t(\Phi_o)$ to denote the projected-then-rotated vector of the corresponding state x_t:

$$\alpha_t(\Phi_o) \quad \triangleq \quad R(\Phi_o)H_{l_t}x_t \tag{27}$$

Combining terms, we obtain the observation log-likelihood :

$$\mathbb{L}^i(\Phi_o) \quad \equiv \quad -\sum_{t=1}^{|Z|} \Big\langle [z_t - \alpha_t(\Phi_o)]^T V_{\hat{l}_t}^{-1} [z_t - \alpha_t(\Phi_o)] \Big\rangle_{f_I^i(x_t)} \tag{28}$$

where we have omitted redundant constant terms. Intuitively, the optimization in (28) is to find an updated dance angle Φ_o^{i+1} which minimizes the sum of the expected Mahalanobis distances between the observations z_t's and the projected-then-rotated states $\alpha_t(\Phi_o)$'s. However, since the non-linearity is involved due to the rotation, there is no analytical solution to this maximization problem. Specifically, Eq. 28 involves quadratic trigonometric terms, e.g., $sin(\Phi_o)^2$. Thus, we conduct 1D gradient ascent on the obtained functions where the increase of the model likelihoods is still guaranteed in the spirit of Generalized EM [57].

5.3 Experimental Results

The experimental results demonstrate that PS-SLDSs provide reliable global parameter quantification capabilities along with improved labeling abilities in comparison to the standard SLDS model.

We conducted experiments with 6 video sequences with length 1058, 1125, 1054, 757, 609 and 814 frames, respectively. Once the sequence observations Z's were obtained, the trajectories were pre-processed so that the mean of each track is located at a fixed coordinate origin Note from Fig. 15 that the tracks are noisy and much more irregular than the idealized stylized dance prototype shown in Fig. 14(a). The red, green and blue colors represent right-turn, waggle and left-turn phases. The ground-truth labels are marked manually for comparison and learning purposes. The dimensionality of the continuous hidden states was set to four.

We adopted a leave-one-out (LOO) strategy for evaluation. The parameters are learned from five out of six datasets, and the learned model is applied to the left-out dataset to perform labeling and simultaneous quantification of angle/average waggle duration. Six experiments were conducted using both PS-SLDSs and standard SLDSs, so that we test on each sequence once. The PS-SLDS estimates of angle and average waggle durations (AWD) are directly obtained from the results of global parameter quantification. On the other hand, the SLDS estimates are heuristically obtained by averaging the transition numbers or averaging the heading angles from the inferred "waggle" segments.

5.3.1 Learning from Training Data

The parameters of both PS-SLDSs and standard SLDSs are learned from the data sequences depicted in Fig. 15. The standard SLDS model parameters were learned as described in the section for learning in SLDSs based on the given ground truth labels. The canonical parameter tuples described in Section 5.2.1 are all learned solely based on the observations Z without any parameter tying. However, the prior distribution π on the first label was set to be a uniform distribution.

To learn the PS-SLDS model parameters, the ground truth waggle angles and AWDs

were manually found from the data. Then, each sequence was pre-processed (rotated to be aligned) in such a way that the waggles head in the same direction based on the estimated ground truth waggle angles. This preprocessing was performed to allow the PS-SLDS model to learn the canonical parameters which represent the behavioral template of the dance. Note that the six sets of model parameters are learned through the LOO approach and the global angle of the test sequence is not known a priori during the test phases. In addition to the model parameters learned by the standard SLDS, PS-SLDSs learn additional duration models D (with focus on the duration variances Σ), and the semi-Markov transition matrix $\tilde{B}$, as described in Section 3.2.1.

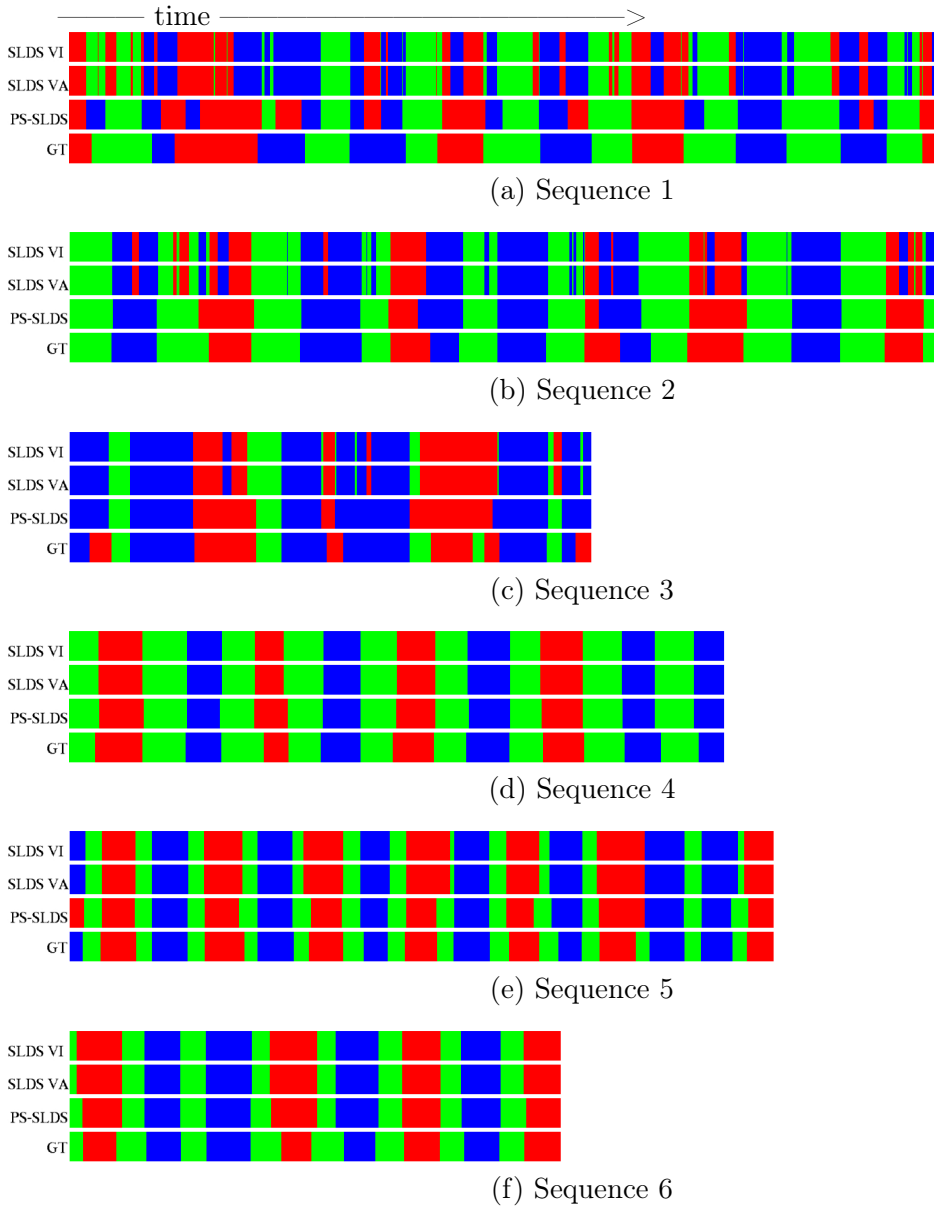
5.3.2 Inference on Test Data

During the testing phase, the learned parameter sets are used to infer the labels of the left-out test sequences. An approximate Viterbi (VI) method [67, 69] and a variational approximation (VA) methods [60] were used to infer the labels in standard SLDSs. The initial probability distributions for the VA method were initialized based on the VI labels. Our initialization scheme assigned VI labels a probability of 0.8 and the other two labels at every time-step were assigned probabilities of 0.1. We used the VI method due to its simplicity and speed. Our experiments compare the performance of SLDS and PS-SLDS models based on VI and VA methods.

5.3.3 Qualitative Results

Our experimental results demonstrate the superior recognition capabilities of the proposed PS-SLDS model over the original SLDS model. The label inference results on all data sequences are shown in Fig. 16. The four color-coded strips in each figure represent SLDS VI, SLDS VA, PS-SLDS VI and the ground-truth labels from the top to the bottom. The x-axis represents time flow and the color is the label at every corresponding video frame.

The superior recognition abilities of PS-SLDSs can be observed from the presented results. The PS-SLDS results are closer to the ground truth or comparable to SLDS results in all sequences. In particular, the sequences 1, 2 and 3 are challenging because the tracking results obtained from the vision-based tracker are more noisy. In addition, the patterns



(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

(e) Sequence 5

(f) Sequence 6

Figure 16: Label inference results. Estimates from SLDS and PS-SLDS models are compared to manually-obtained ground truth (GT) labels.
Key : waggle, right-turn, left-turn

Table 2: Absolute errors in the global rotation angle estimates from PS-SLDS and SLDS in radians. The numbers in parenthesis are error rates (%). Last row contains the ground truth rotation angles. Sequence numbers refer to Fig. 15.

Sequence	1	2	3	4	5	6
PS-SLDS	0.09 (30)	**0.01** (4)	0.03 (3)	0.11 (8)	**0.11** (5)	**0.06** (8)
SLDS VI	**0.05** (16)	0.03 (12)	**0.02** (2)	**0.09** (7)	0.18 (9)	0.09 (11)
SLDS VA	**0.05** (16)	0.03 (12)	**0.02** (2)	**0.09** (7)	0.18 (9)	0.09 (11)
Ground Truth	-0.30	-0.25	1.13	-1.33	-2.08	-0.80

Table 3: Absolute errors in the Average Waggle Duration (AWD) estimates for PS-SLDS and SLDS in frames. The numbers in parenthesis are error rates (%). Last row contains the ground truth AWD. Sequence numbers refer to Fig. 15.

Sequence	1	2	3	4	5	6
PS-SLDS	**13.7** (27)	**0.91** (2)	**1.9** (9)	**0.22** (<1)	**0.4** (2)	**5.6** (17)
SLDS VI	40.8 (79)	28.9 (62)	11.1 (52)	0.44 (1)	3.6 (19)	8 (25)
SLDS VA	40.7 (79)	28.9 (62)	11.1 (52)	0.44 (1)	3.6 (19)	8 (25)
Ground Truth	51.6	46.6	21.4	41.1	19.4	32.6

Table 4: Accuracy of label inference in percentage. Sequence numbers refer to Fig. 15.

Sequence	1	2	3	4	5	6
PS-SLDS	**75.9**	**92.4**	**83.1**	**93.4**	**90.4**	**91.0**
SLDS DD-MCMC	74.0	86.1	81.3	**93.4**	90.2	90.4
SLDS VI	71.6	82.9	78.9	92.9	89.7	89.2
SLDS VA	71.6	82.8	78.9	92.9	89.7	89.2

of switching between the dance modes and the durations of the dance regimes are more irregular than the other sequences.

It can be observed that most of the over-segmentations that appear in the SLDS labeling results disappear in the PS-SLDS labeling results. PS-SLDSs labels still introduce some errors, especially in sequences 1 and 3. However, keeping in mind that even a human expert can introduce labeling noise, the labeling capabilities of PS-SLDSs are fairly good.

5.3.4 Quantitative Results

The quantitative results on the angle and average waggle duration quantification show the robust global parameter quantification capabilities of PS-SLDS. Table. 2 shows (from top to bottom) : the absolute errors of the PS-SLDS estimates along with the error rates (%)[3]

[3]The error rates are obtained by dividing each estimate by the corresponding ground truth value.

in parenthesis, SLDS estimates based on the VI and the VA methods, and the ground truth angle. The best estimates are accented in bold fonts. The SLDS estimates for the global parameters are obtained by the heuristic of averaging the heading angles in the sequences that were labeled as "waggle" in the inference step. All of the error values are the difference between the estimated results and known ground truth values.

Based on the six tests, PS-SLDSs and SLDSs show comparable waggle angle estimation capabilities. There is no distinguishable gap in performance between VI and VA methods. Our hypothesis is that the over-segmentation errors do not effect the waggle angle estimates as much as it effects the average waggle duration estimates, since the waggle segments detected by SLDSs are still mostly correct in spite of the over-segmentation effects. Note that the maximum error of PS-SLDS angle estimate was 0.11 radians for the fifth dataset, which is fairly good considering the noise in the tracking results.

The quantitative results on average waggle duration (AWD) quantification show the advantages of PS-SLDS in quantifying the global dynamics parameters of interest. AWD is an indicator of the distance to the food source from the hive and is a valuable data for insect biologists. Table. 3 shows (from top to the bottom) : the absolute errors and error rates of the PS-SLDS estimates, the SLDS estimates of VI and VA methods and the ground truth AWDs. Again, the best estimates are marked in bold fonts where PS-SLDS estimates are consistently and substantially superior to the SLDS estimates. The SLDS estimates are obtained by evaluating the means of the waggle durations in the inferred segments. The results again show that PS-SLDS estimates match the ground-truth closely. In particular, we would like to highlight the quality of the PS-SLDS AWD estimates for sequences 2, 3, 4 and 5. In contrast, the SLDS estimates in these cases are inaccurate. More specifically, the SLDS estimates deviate far from the ground truth in most cases except for the sequence 4. The reliability of the AWD estimates obtained by PS-SLDSs show the benefit of the duration modeling and the canonical parameters supported by the enhanced models.

Finally, Table 4 shows the overall accuracy of the inferred labels for the PS-SLDS, SLDS

DD-MCMC [61][4], SLDS VI, and SLDS VA results. It can be observed that PS-SLDS provides very accurate labeling results w.r.t. the ground truth. Moreover, PS-SLDS consistently improves upon the standard SLDSs across all six datasets. The overall experimental results show that PS-SLDS model is promising and provides a robust framework for the bee application. It should be noted that SLDS DD-MCMC is the most computationally intensive method, and PS-SLDS still improves on SLDS DD-MCMC in a consistent manner.

5.4 Conclusion

We presented experimental results on the real-world honey bee dance sequences, where the honey bee dances were modeled using PS-SLDSs. Both the qualitative and quantitative results demonstrate that the enhanced PS-SLDS model can robustly infer the labels and global parameters more accurately in comparison to SLDSs. Accordingly, the accurate quantification abilities of PS-SLDSs validate the additional modeling efforts to include global parameters on top of the simpler base models. The consistently superior results obtained by PS-SLDSs for the honey bee dance data suggest that PS-SLDSs may be promising for other applications.

[4]The data-driven Markov chain Monte Carlo (DD-MCMC) inference method for SLDSs [61] was developed by the author, to examine the full potential capabilities of SLDSs with least approximation.

Chapter VI

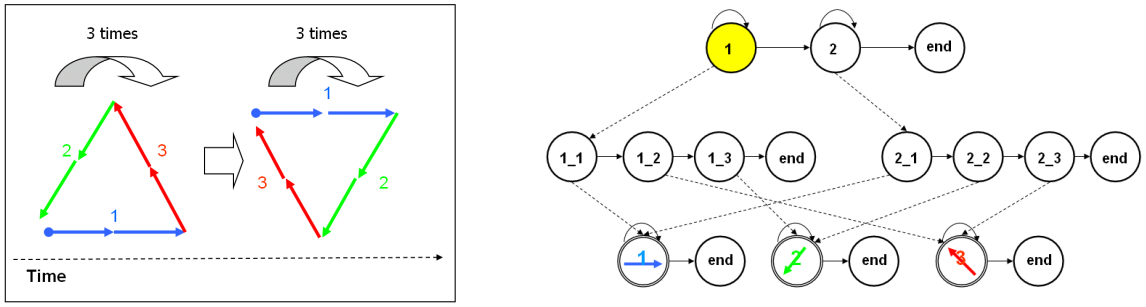
HIERARCHICAL SWITCHING LINEAR DYNAMIC SYSTEMS

6.1 Introduction

In this chapter, we introduce *hierarchical switching linear dynamic systems* (H-SLDSs), a probabilistic time-series model designed to encode the dynamics and the hierarchical temporal ordering structure exhibited by multivariate signals.

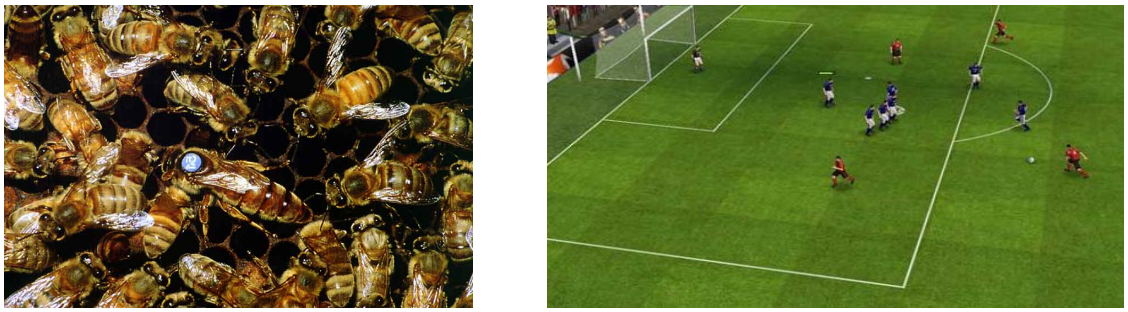
The superior aspect of the hierarchical models over flat temporal models is the ability to reuse sub-models in different contexts, which can be observed well in the following example. A hierarchical automaton model that corresponds to the upward-downward triangle sequence in Fig. 17(a) is illustrated in Fig. 17(b). It can be observed that the shared three primitive dynamic patterns (colored in blue, red, and green) appear in two different upward and downward triangle sub-structures in the sequences. In the flat Markov models, such repeating sub-structures (in the middle layer) or primitive dynamic patterns (at the bottom level) need to be redundantly duplicated under the left-to-right modeling assumption.

Other examples of real-world temporal data which exhibit intrinsic hierarchy are honey bees and soccer players, shown in Fig. 18. The first example, the honey bee community, is shown in Fig. 18(a). The honey bee community consists of three different types of members,



(a) (b)

Figure 17: (a) An example upward-downward triangle sequence. (b) An example 3-level hierarchical automaton representing the triangle sequence. Solid lines represent horizontal transitions, dotted lines represent vertical transitions. Double-circled nodes represent production states on which the corresponding dynamic patterns are visually overlaid.



(a) (b)

Figure 18: (a) A scene of honey bee hive : a queen bee is color-marked in the middle, surrounded by drones and worker bees. (b) A shot of a soccer game where each team consists of multiple players with different roles.

namely a queen, worker bees and drones. While the low-level primitive motion patterns of all three types of bees over short time duration would be rather similar, it is the clear difference in the longer-term temporal patterns that lets us to identify the roles of each bee and the collective status of the community from their motion trajectories : a queen bee has relatively less dynamic motion range over time, while worker bees have the largest range of motion such as dancing within the hive and staying on and off from the hive. Another example is a soccer game, shown in Fig. 18(b). A soccer team consists of multiple players whose roles differ substantially depending on their positions, i.e., attack middle fielders, defenders, a goal keeper, center forwards, and etc. Again, the motion trajectories of each player over short duration do not provide strong cues on the players' roles and the current team strategies. However, the longer-term trajectories provide us with relatively strong clues to answer many different types of high-level questions. Hence, a hierarchical model can be used to identify the top-level role of every player as well as the strategies the players are following. Accordingly, the hierarchical models can be used to label the play of every player at multiple semantic/temporal resolutions.

The real-world data of interest in this dissertation for the hierarchical modeling work is the human dumbbell exercises conducted in the gyms. For example, most gym exercises can be grouped into either aerobics/an-aerobics activities. Then, they can be further categorized down-to upper/lower/whole body exercises. Eventually, there are many different types of weight-lifting or dumbbell exercises which share common low-level motions such as bend,

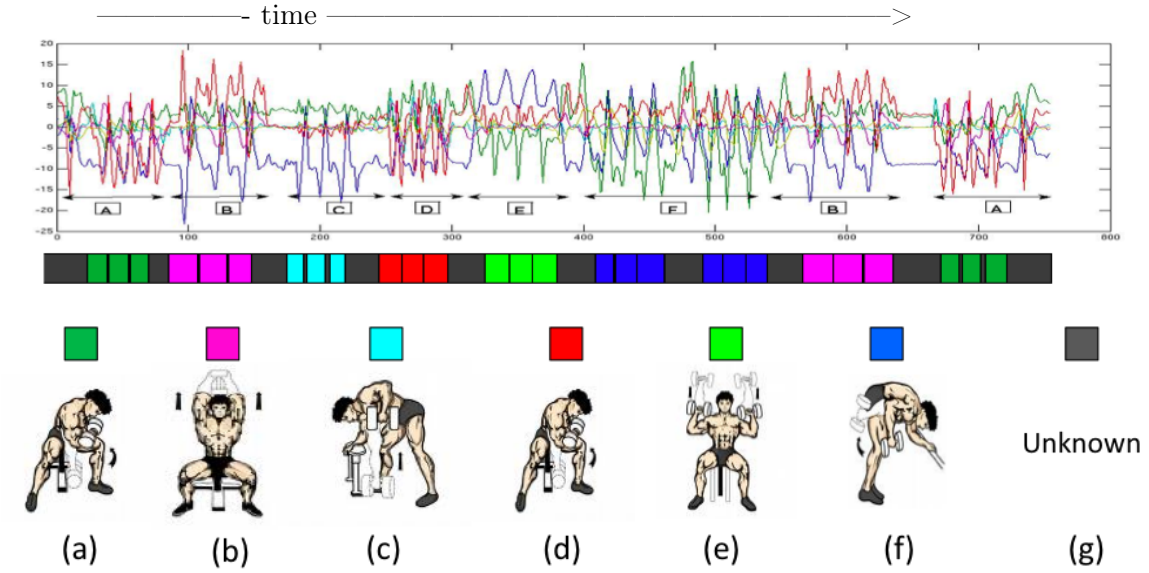


Figure 19: The top figure shows the six dimensional signals collected from a wired on-body sensor where the subject conducted six (seven including unknown) different dumbbell exercises illustrated at the bottom : (a) flat curl, (b) shoulder extension, (c) back, (d) twist curl, (e) shoulder press, (f) tricep, and (g) unknown. Every occurrence of the exercises is visualized as a colored rectangle where the labels are shown as a color strip below the top figure with the color and the width of the rectangles corresponding to the category and the duration of conducted exercises.

twist, extend, curl and etc. As an example, Fig. 19 shows a six dimensional signal sequence collected from a wired on-body sensor where the subject conducted six different dumbbell exercises illustrated at the bottom. It can be observed that the raw signals demonstrate bounded number of patterns repeatedly which are shared across different exercise categories.

The development of H-SLDSs presented in this paper has been motivated to achieve the following goals. First, a hierarchical extension of standard SLDSs [10, 69] which can encode the interaction between the temporal structures at different granularities was needed. Second, a scalable model representation whose size increases moderately w.r.t. the size of the problem was necessitated. Third, a hierarchical formulation which allows us to interpret novel data simultaneously at multiple temporal resolutions was needed. Finally, we aimed to understand how the resulting theoretical framework solves large-sized real-world problems.

The previous work that partly addressed the hierarchical extensions of SLDSs include 'Dynamical System Trees' (DSTs) [37][1] and 'Multi-scale SLDSs' (MS-SLDSs) [91], whose

[1]A DST has a tree structure and is targeted to model a group of interacting temporal sequences. The

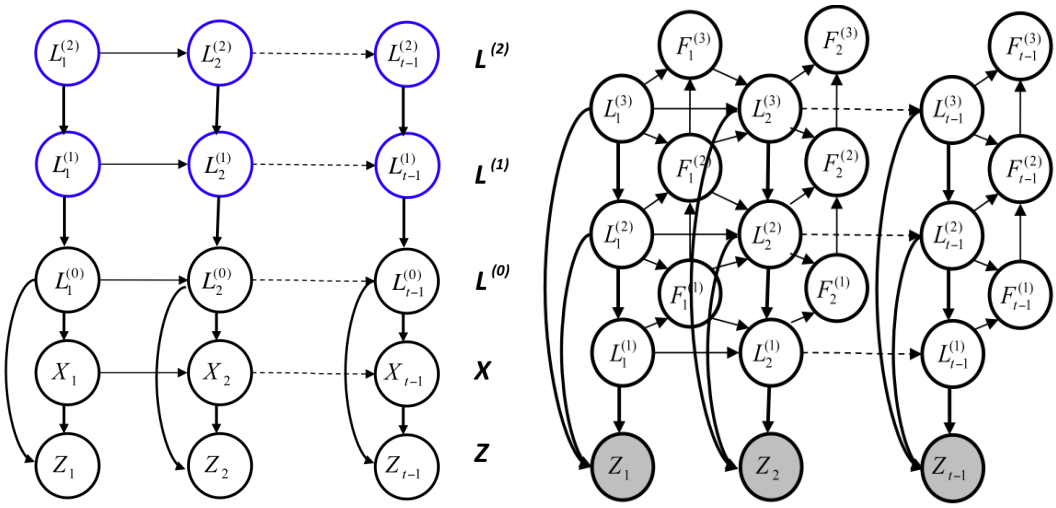


Figure 20: Dynamic Bayesian networks of related work. (a) A hierarchical extension of SLDSs used in [91, 37] with the hierarchical Markov chain colored in blue. (b) A hierarchical HMM model presented in [56].

graphical model is shown in Fig. 20(a). These work extended the flat SLDSs by introducing multi-layer hierarchical Markov chains. However, the representational power of these models were limited in the sense that they do not provide a principled mechanism to encode terminal sub-states for the parent states. Hence, the parent states can switch at anytime regardless of the configuration of the child states and the models are not able to encode the call-return semantics between the hierarchies. Accordingly, DSTs and MS-SLDSs are not able to describe data more descriptively even in the presence of well-established domain knowledge about terminal sub-states and such shortcomings may potentially undermine the accuracy of the inference tasks.

A previous work that provided a solution to represent the terminal sub-states and termination probabilities within the graphical model formalism was [56] where the corresponding DBN is shown in 20(b). They explicitly added 'finish' variables (marked as F variables) within the proposed hierarchical DBN which encode the 'return' semantics of the terminal sub-states. The H-SLDS model to be presented in the following sections share similar architecture to H-HHMs in that H-SLDSs adopt hierarchical Markov chain with the finish

DBN shown in Fig. 20(a) corresponds to a branch of a DST.

variables to encode the higher-order structure within data, with the necessary modifications suitable to encode continuous signal domains that SLDSs address.

In terms of scalability, a promising avenue has been shown for the visual object recognition task [83] where the sharing of the sub-structures was proposed as a solution. It suggests that, by sharing the sub-states between the upper-level states, the size of the model may increase sub-linearly w.r.t. the number of the new states added at a higher level by re-using most of the existing parameters for the sub-states. Although a similar motivation of sub-structure sharing was mentioned in an H-HHM work [56], the described model had a dense inter-dependency between all the layers which made it unclear how the sub-structures should be shared. Another work on H-HHMs which also incorporates finish variables within the hierarchical Markov chain additionally address the sub-structure sharing problem [17], but presents an inference algorithm with the computational complexity $O(T^3)$. Because we are interested in the data collected over a long period of time, the implied cubic computational complexity would not be suitable for our purpose.

Another potential advantage that we can gain by relying on hierarchical models is the computational efficiency or saving. In detail, the shared sub-structures provide us with the possibility to re-use a large amount of the computational results, resulting in the speed up of inference. Especially, this is advantageous in the case where we are interested in the labeling tasks. The demanding computations associated with the filtering or the smoothing operations for the LDS primitives at the bottom level can be effectively shared under different contexts of high-level state configurations.

Our work on H-SLDSs presents the following contributions. First, the developed H-SLDS model provides superior descriptiveness in modeling over the previous hierarchical extensions of SLDSs [91, 37], by introducing the finish variables which encode the conditions when every Markov chains would terminate. Second, we provide novel sub-structure representations and the associated learning methods for each which allow us to effectively share sub-structures within a hierarchy to achieve scalability, which are three-fold : (1) we present an approach to discover a set of base LDS vocabulary to be shared, by learning a mixture of LDSs from the post-processed data, (2) left-to-right SLDSs (LR-SLDSs), which impose minimal

assumptions on the Markov transitions between LDSs, are presented to provide the upper-level representations which are assembled from the learned LDS vocabulary, (3) the inter-dependence between the layers in the hierarchy is designed to be sparse to enhance the modular behavior between hierarchies. Third, we provide effective methodology for inference in H-SLDSs where we convert H-SLDSs to equivalent SLDSs to use the existing array of inference methods for SLDSs. Finally, we provide a strong empirical proof on the usefulness of the above H-SLDSs and hierarchical temporal models in general, by applying the model to the two real-world human exercise datasets whose hierarchy structures are carefully designed to possess interesting temporal structure and the accuracy of labeling tasks are measured against the ground truth labels. More in detail, for the thoroughness of the empirical analysis, the inference results obtained using H-SLDSs were compared to the ground truth labels across the entire hierarchy to provide both qualitative and quantitative analysis of the results. We believe that the empirical analysis presented in this thesis provides stronger evidence on the usefulness of the hierarchical temporal models than the previous work where only the limited toy examples [91, 37] were tested or the behavior of the model could not be analyzed in-depth since the ground-truth did not exist [37].

In the remainder of the chapter, we will describe the sub-structure models and the methodologies that are used to construct H-SLDSs : (1) the learning methodology to learn LDS vocabulary to be shared across distinct Markov chains, and (2) the left-to-right SLDS models (LR-SLDSs) which are used to encode the lowest-level temporal patterns, which correspond to the leaf states of the hierarchical Markov chains. Finally, we will present the graphical model of H-SLDSs with the details on the joint PDFs along with the appropriate inference methodologies.

6.2 Learning Shared LDS Vocabulary

The approach adopted in this work to build a scalable hierarchical extension of SLDSs is to promote the sharing of the bottom-level LDSs. For example, a set of low-level motions, e.g., twist, untwist, bend, extend, and etc., are shared between different weight-lifting or dumbbell exercises illustrated in Fig. 19. It would be reasonable to expect that the representational

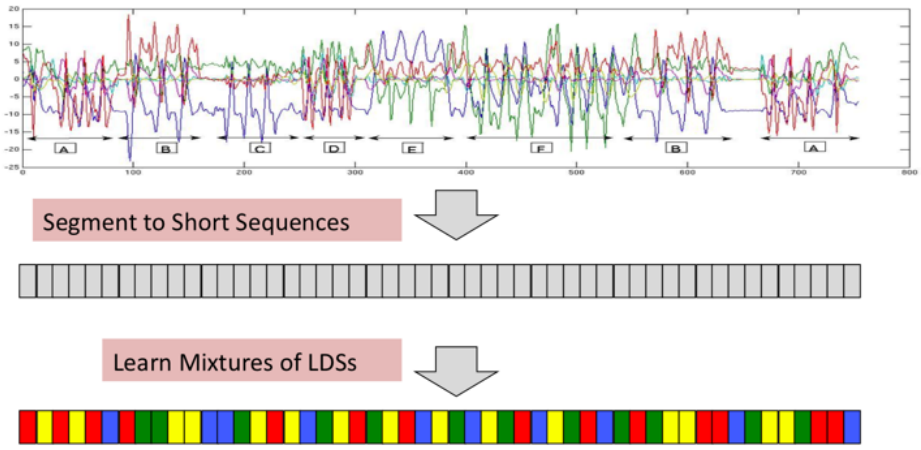


Figure 21: An illustration of the LDS vocabulary learning scheme. A sequence of six dimensional data shown at the top is chopped into a set of unlabeled short segments. Then, a set of N LDS models are learned via clustering within EM framework where each segment will be assigned the most likely cluster membership and the LDS models are learned.

size of the agglomerate model would be much smaller when such low-level dynamic primitives are discovered and shared between multiple high-level exercise categories.

To discover a set of common LDS primitives embedded in the signals, we take a machine-learning approach where a mixture of LDSs are learned from the collected signals once the signals are chopped into short segments of fixed duration. For example, a sequence of six dimensional data (30 Hz) collected for about 1.5 minutes shown at the top of Fig. 21 is chopped into a set of unlabeled short segments, as shown in the middle of Fig. 21, we can learn a set of N LDS models by clustering the segments based on their likelihood within EM framework where each short segment will be assigned the most likely cluster membership, as visualized in different colors at the bottom of Fig. 21. Simultaneously, the LDS parameters are learned from the (probabilistically) assigned segments. Afterwards, the parameters of LDSs can be possibly improved by building an SLDS model from them, and updating the parameters through further learning processes. However, we did not see any benefits of such further learning process in our application domain described in Chapter 7.

In practice, we have set the duration of short segments to be physically sensible, a half second each for human gym exercises. Then, a mixture of fifteen LDS models are learned, based on a variant of an EM algorithm presented in [20]. The number of LDS vocabularies,

e.g., $N = 15$ for the gym exercise data, can be chosen empirically to be the maximum size where every LDS model represents at least a minimum percentage, e.g., 2 percent, of the whole segments. More principled model selection method such as AIC [3] can be used but overly simplified models were observed to be preferred, and such options were not investigated further.

6.3 Left-to-right SLDSs (LR-SLDSs)

In this section, we present left-to-right SLDSs (LR-SLDSs) which are the sub-structure representations used to encode the lowest-level semantic temporal patterns that form the leaf states of the hierarchical Markov chain within H-SLDSs. The graphical model of LR-SLDSs is shown in Fig. 23(a) where the left-to-right Markov chain is colored in blue. The examples of the lowest-level semantic temporal patterns would correspond to exercise categories such as flat-curl or shoulder extension in our wearable exercise dataset. In other words, the lowest-level semantic patterns correspond to labels at the finest level whose examples can be manually provided by human experts.

The LR-SLDS model provides the following desirable properties :

- The top chain of LR-SLDSs (shown in blue in Fig. 23(a)) is a well-known left-to-right Markov chain [13] where the left-most and the right-most states correspond to the start and end characteristics of the temporal signals. Hence, LR-SLDSs are suitable to capture the one-way ordering structures within temporal patterns.

- Every state within the left-to-right Markov chain is associated with emission model over the learned LDS vocabulary. Hence, LR-SLDSs builds implicit transition models between LDSs rather than adopting more explicit Markov transition models used in standard SLDSs.

The rationale to use the LR-SLDS model over the standard SLDS model is that it allows us to re-use shared LDSs with minimal assumption on the Markov transition model, which attempts to soothe the encoding of direct temporal correlations between LDSs. An alternative but common choice for the discrete Markov chain structure would be the fully connected

model between all the LDSs. However, such models may be problematic in case (1) identical LDSs may appear multiple times in a pattern where the learned model would be less descriptive due to the averaging effect, or (2) the level of the noise in the raw signals Z are high enough to create the risk of over-fitting where the learned model would be misleading under the stronger assumption of fully connection Markov transition model. Additionally, LR-SLDSs are intended to be used to encode the temporal patterns at the bottom hierarchy of H-SLDSs where the risk of failing to capture the repeating patterns are minimized. Above the LR-SLDS layer, LR-SLDSs are actively shared across hierarchical Markov chain (as described in Section 6.4 and LR-SLDSs build upon the learned LDS vocabulary. Hence, the use of LR-SLDSs as our intermediate models is exactly in line with the overall goal of actively sharing sub-structures across the hierarchical models.

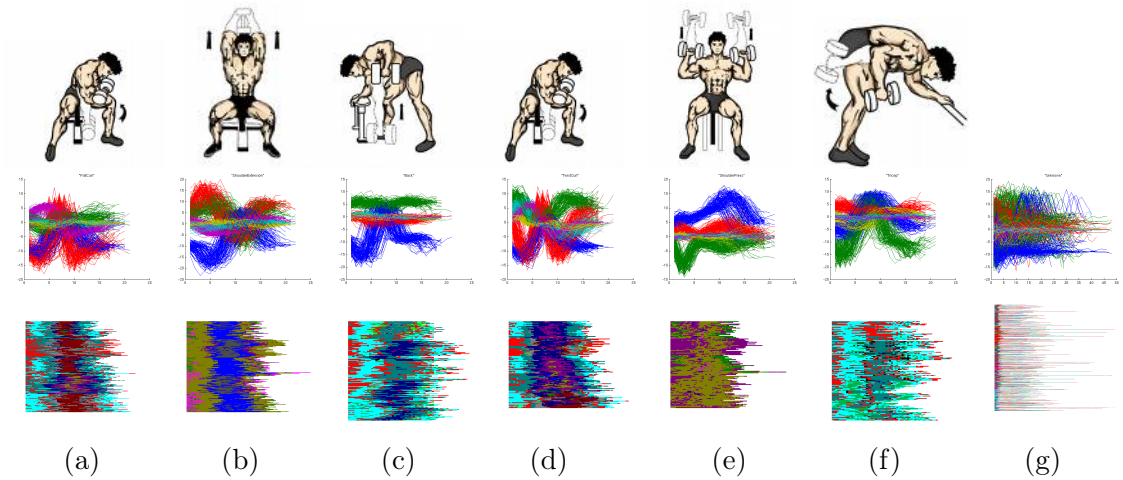
In detail, the top chain of LR-SLDSs (shown in blue in Fig. 23(a)) is a well-known left-to-right Markov chain denoted by $L^{(1)}$ where the super-script denotes the height of the discrete layer within the entire hierarchy, with the zeroth layer corresponding to the LDS vocabulary. During the generative process, the left-to-right chain initializes w.r.t. a prior distribution π, i.e., $l_1^{(1)} \sim \pi(l_1^{(1)})$. Then, as the state $l_t^{(1)}$ at the top chain evolves over time based on a Markov transition model $B \triangleq P(l_t^{(1)}|l_{t-1}^{(1)})$, the state $l_t^{(1)}$ emits an LDS primitive $l_t^{(0)}$ among a set of n learned LDSs, based on an emission model $E_{l_t} \triangleq P(l_t^{(0)}|l_t^{(1)})$ where $E \triangleq \{E_i|1 \leq i \leq n\}$. Hence, the joint PDF of LR-SLDSs is as follows :

$$\begin{aligned} P(L^{(1)}L^{(0)}XZ) &= P(XZ|L^{(0)})P(L^{(0)}|L^{(1)})P(L^{(1)}) \\ &= P(XZ|L^{(0)}) \left\{ \prod_{t=1}^{T} P(l_t^{(0)}|l_t^{(1)}) \right\} P(L^{(1)}) \end{aligned}$$

In summary, an LR-SLDS model is completely defined by a parameter set $\Phi \triangleq \{\pi, B, E, M\}$ where M denotes a set of LDS model parameters.

6.3.1 Inference and Learning in LR-SLDSs

In this work, the inference and learning in LR-SLDSs are conducted in an analogous way to the methods used for S-SLDSs in Ch. 3, i.e., model conversion. All the possible pairs of left-to-right Markov chain states and LDS states are formed to create meta states. Consequently,



(a) (b) (c) (d) (e) (f) (g)

Figure 22: Each figure corresponds to the following category : (a) flat curl (b) shoulder extension (c) back (d) twist curl (e) shoulder press (f) tricep (g) unknown. The top row images show the snapshot of each exercise pattern. The images in the second row shows the multivariate time series data in each category. The bottom row shows the labeling results of data in each category using the corresponding learned LR-SLDS model where the LDS labels are color-coded.

an LR-SLDS model is converted to an equivalent SLDS model where an approximate inference method for SLDSs is used to conduct inference. Finally, the inference results are marginalized to obtain the results for LR-SLDSs. For learning, the parameters of LR-SLDSs can be learned within the EM framework, as has been done for S-SLDSs.

In terms of initialization, the left-to-right Markov transition model can be initialized to be a band-diagonal upper diagonal matrix where the rows and columns correspond to the previous and the next states respectively. The emission models can be initialized in a data-driven manner. In detail, the whole training sequences can be heuristically labeled by classifying every time frame along with its neighborhood deterministically w.r.t. the learned LDS models, by assigning the LDS label which incurs highest likelihood for that particular time frame. Once such labels are obtained, the whole label sequences are divided into segments with equal lengths. Then, each emission model can be initialized based on the empirical label distribution that appears in the corresponding segments on top of the uninformative uniform prior.

We used the LR-SLDS models to encode the dumbbell exercise data where there were total 7 categories which include 6 distinct dumbbell exercises and an additional unknown

category. First, sets of training data for each category were collected manually where LR-SLDSs are to build upon the shared 15 LDS models. Illustrative figures of exercises are shown at the top row of Fig. 22. The corresponding six-dimensional data sequences, overlaid on top of each other, are shown in the middle row for each category, demonstrating the noticeable within-category similarity in their signals. Note that the different dimensions within signals are visualized with distinct colors to illustrate the behavior of the signals more descriptively. Then, LR-SLDSs were initialized with Markov transition model with bandwidth of two where such bandwidth was chosen to descriptively capture the temporal ordering structure while allowing a plausible amount of variation. Finally, the LR-SLDS models are learned through EM method using approximate Viterbi method as the inference tool during the E-steps, as described earlier. The switching behavior between LDSs interpreted by LR-SLDSs are shown at the bottom row of Fig. 22 where the different LDS components are visualized with distinct colors. In other words, the color strips at the bottom row are the LDS labels assigned to the $L^{(0)}$ layer in Fig. 23(a). The horizontal and vertical axes correspond to time and sequences respectively. It can be observed that the signals in each category demonstrate similar temporal ordering structure while variations do exist. Moreover, the temporal ordering structure of (a) flat-curl and (d) twist-curl show very similar patterns where the only real-world difference in their motion is the wrist twist that occurs in the middle of the trajectories. It seems that such subtle difference in the dynamic patterns are successfully captured by the LR-SLDS models to a certain extent.

In the next section, we describe the details of H-SLDSs and show how they incorporate hierarchical Markov chains on top of LR-SLDSs to encode longer-term temporal structures of higher-level regimes.

6.4 Hierarchical SLDSs

We introduce H-SLDSs, a novel hierarchical extension of SLDSs. The H-SLDS model is formulated by incorporating the hierarchical Markov chains augmented by finish variables on top of a set of LR-SLDSs which are built from a set of shared LDSs. In the following sections, we describe the parameterization of H-SLDSs in detail along with its potential use

for the targeted wearable exercise interpretation tasks.

6.4.1 Graphical model of H-SLDSs

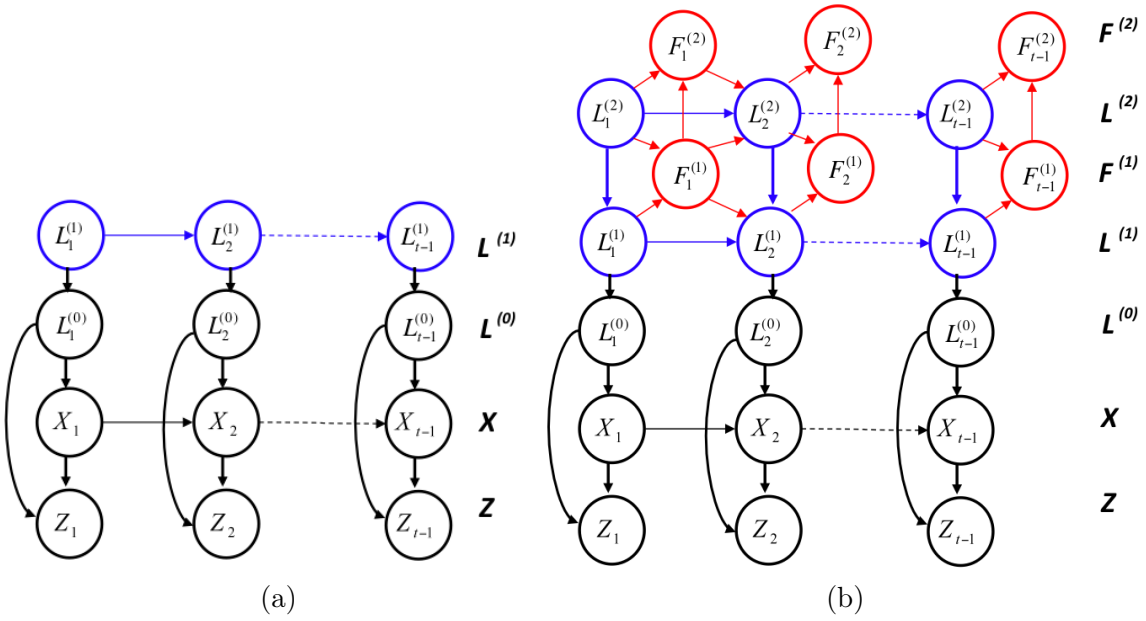


Figure 23: Dynamic Bayesian networks of LR-SLDSs and H-SLDSs. (a) A left-to-right SLDS model. The blue sub-structure denotes the left-to-right Markov chain. (b) An H-SLDS model with two level hierarchy. The blue and red sub-structures correspond to the hierarchical Markov chain and the finish variables respectively. The bottom Markov chain at $L^{(1)}$ layer corresponds to the left-to-right Markov chain. The discrete nodes underneath, denoted by $L^{(0)}$, correspond to the emitted LDS dynamic modes. It can be observed that H-SLDSs build upon LR-SLDSs by introducing additional hierarchical Markov chain at the upper levels.

The DBN for the H-SLDS model with hierarchical Markov chains (HMC) is shown in Fig. 23(b) where the example has two-layer Markov chain marked in blue. Each discrete node in the DBN is denoted as $l_t^{(h)}$ with the superscript h corresponding to the height of the layer (base zero) where the bottom level $L^{(0)}$, which is not a part of the HMC, corresponds to the activated LDSs. A time-slice of the DBN at time t for H-SLDSs comprises of the discrete state vector $L_t^{0:H} \triangleq [l_t^0, l_t^1, ..., l_t^H]$ where the height of the Markov chain is H, the continuous state x_t, and the corresponding observation z_t.

The H-SLDS model builds upon LR-SLDSs by introducing two additional structures : (1) the (blue) HMC with the downward arcs from the upper layers to the lower layers, which encode the Markov behavior of the lower-level sub-states in the context of the parent

states, (2) the (red) Boolean finish variables $f_t^{(h)}$ are designed to identify the lower-level sub-states which trigger the termination of the current Markov regime and the transition at the parent level. In particular, the diagonal upward arcs from the finish variables $f_t^{(h)}$ at the h-th layer to the discrete states $l_{t+1}^{(h+1)}$ are designed to identify the lower-level states which trigger the state transitions at the upper levels. Simultaneously, the finish variable structure helps the inference phase to decode the label (segmentation) boundaries more accurately and to enforce the property that transitions in the upper layers occur only when there is a transition in the lower level. A related hierarchical SLDS extension without the finish variables appeared in [37, 91]. Hence, the presented H-SLDS model in Fig. 23(b) is novel primarily due to the newly added finish variables and the use of LR-SLDSs as its core sub-structure representations.

More in detail, the downward arcs between the discrete states $l_t^{(h+1)}$ and $l_t^{(h)}$ represent the fact that a parent state invokes child sub-states. Additionally, the finish variable $f_t^{(h)}$ at the h-layer is a Boolean variable which indicates that the current Markov regime conditioned by the parent state $l_t^{(h+1)}$ finishes at time t when $f_t^{(h)} = true$. In the case where $f_t^{(h)} = false$, the current regime at the h-th layer continues with the the constant ancestor configuration remaining constant, i.e., $L_{t+1}^{(h+1):H} \leftarrow L_t^{(h+1):H}$. Note that if $f_t^{(h)} = true$, then for all the levels below $h' < h$, the finish variables are $true$, i.e., $\forall h' < h$, $f_t^{(h')} = true$. In other words, a parent state invokes the Markov regime which guides the switching patterns of the child sub-states by the downward arrows. Then, once such Markov regime terminates, the finish variables turn on and signals back to the parent state (via the upward diagonal arcs) to invoke a state transition.Additionally, the (red-colored) upward arcs between the f variables across different levels demonstrate the fact that a parent regime can only switch when the lower sub-regime of the child states finishes, effectively enforcing call-return semantics.

The main difference between the structure of the HMC presented in this work and the work in [56] is that most of the incoming arrows from the ancestor variables $L_t^{(h+1):H}$ to every discrete state $l_t^{(h)}$ disappear and only a single incoming downward arrow from its parent state l_t^{h+1} exists in our work. The sparser structure implies that the behavior of a particular layer depends only on the immediate parent and child layers, enforcing an even

more modular call-return semantics.

6.4.2 Conditional PDFs of HMC

In H-SLDSs, the discrete state $l_t^{(1)}$ at the bottom HMC layer is evolved from the previous state $l_{t-1}^{(1)}$ depending on the associated finish variable $f_{t-1}^{(1)}$ and its parent state $l_t^{(2)}$. In the case where the finish variable was turned off, i.e., $f_{t-1}^{(1)} = false$, the state $l_t^{(1)}$ maintains its regime and switches based on the left-to-right Markov transition model $B^{l_t^{(2)}}$ of the upper state $l_t^{(2)}$, which indicates the active LR-SLDS model. Otherwise, the finish variable is on, i.e., $f_{t-1}^{(1)} = true$, and it implies that the current LR-SLDS regime on the HMC bottom layer is over. Hence, the new state $l_t^{(1)}$ is sampled from the prior multinomial distribution $\pi^{L_t^{(2)}}$ of the new LR-SLDS parent state $l_t^{(2)}$. Formally, we can write this as follows where the prior and the transition model that belong to the parent state k are denoted by π^k and B^k and exclamation marks (!) denote negations :

$$P(l_t^{(1)} = j | l_{t-1}^{(1)} = i, l_t^{(2)} = k, f_{t-1}^{(1)} = v) = \begin{cases} \pi^k(j) & \text{if } v \\ B^k(i,j) & \text{if } !v \end{cases} \tag{29}$$

$$= \left(B^k(i,j)\right)^{1-v} \times \left(\pi^k(j)\right)^v \tag{30}$$

In the intermediate h-th layer ($h > 1$), the new discrete state $l_t^{(h)}$ depends on the previous state $l_{t-1}^{(h)}$, the parent state $l_t^{(h+1)}$, and both finish variables at the same level $f_{t-1}^{(h)}$ and at the child level $f_{t-1}^{(h-1)}$. In the case where the finish variable of the child layer is off, i.e., $f_{t-1}^{(h-1)} = false$, it indicates that the current regime is still active, hence $l_t^{(h)} \leftarrow l_{t-1}^{(h)}$. Otherwise, the finish variable of the child layer is on, i.e., $f_{t-1}^{(h-1)} = true$, and indicates that the current state is switching from the previous state, i.e., $l_t^{(h)} \neq l_{t-1}^{(h)}$. There are two cases depending on whether $f_{t-1}^{(h)}$ is $false$ or $true$. In the first case where $f_{t-1}^{(h)} = false$, it states that the regime for $l_{t-1}^{(h)}$ is still active, i.e., $l_t^{(h+1)} = l_{t-1}^{(h+1)}$. Hence, the new state is sampled based on the Markov transition model of the parent state $l_t^{(h+1)}$. On the other hand, in the case where $f_{t-1}^{(h)} = true$, it indicates that $l_t^{(h+1)} \neq l_{t-1}^{(h+1)}$. Consequently, the new state is sampled based on the multinomial prior distribution of the parent state. Formally, it can be written as follows :

$$P(l_t^{(h)} = j | l_{t-1}^{(h)} = i, l_t^{(h+1)} = k, f_{t-1}^{(h-1)} = b, f_{t-1}^{(h)} = v) = \begin{cases} \delta(j,i) & \text{if } !b \\ B^k(i,j) & \text{if } b\&!v \\ \pi^k(j) & \text{if } b\&v \end{cases} \quad (31)$$

$$= (\delta(j,i))^{1-b} \times \left(B^k(i,j)\right)^{b(1-v)} \times \left(\pi^k(j)\right)^{bv}$$

The finish variable at the h-th layer $f_t^{(h)}$ can be true only when $f_t^{(h-1)} = true$ and the current state $l_t^{(h)}$ is allowed to to terminate where $E^k(l_t^{(h)})$ indicates the probability that the child state $l_t^{(h)}$ would invoke the termination of the parent regime :

$$P(f_t^{(h)} = true | l_t^{(h)} = i, l_t^{(h+1)} = k, f_{t-1}^{(h-1)} = b) = \begin{cases} 0 & \text{if } !b \\ E^k(i) & \text{if } b \end{cases} \quad (32)$$

$$= \left(E^{(k}(i)\right)^{f_t^{(h)} \times b} \times \left(1 - E^{(k}(i)\right)^{(1-f_t^{(h)}) \times b} \times (1 - f_t^{(h)})^{1-b}$$

At the zeroth $L^{(0)}$ layer and the continuous state layer X, the generative process follows the standard SLDS model described in Ch. 2. Finally, for the first time-slice, the multinomial prior for the top level state is defined, i.e., $l_1^{(N)} \sim P(l_1^{(N)})$, and the initialization priors $P(l_1^{(h)} | l_1^{(h+1)})$'s are defined for every pair of adjacent layers for all the layers h's.

6.4.3 Joint PDF of H-SLDSs

In this section, we summarize the joint PDF of H-SLDSs, which includes all of the factors described in the preceding sections : SLDSs, LR-SLDS, and HMC. Briefly, the joint PDF of H-SLDSs is defined as follows :

$$P(L^{1:H}, F^{1:H}, L^0, X, Z) = \underbrace{P(L^{1:H}, F^{1:H})}_{\text{HMC}} \times \underbrace{P(L^{(0)} | L^{(1)})}_{\text{LR-SLDS emissions}} \times \underbrace{P(Z | X, L^{(0)}) P(X | L^{(0)})}_{\text{SLDS}}$$

where the first factor of HMC is defined as belows :

$$P(L^{1:H}, F^{1:H}) = \underbrace{P(L_1^{1:H}) \times P(F^{1:H}|L_1^{1:H})}_{\text{Prior at time 1}} \times \underbrace{\prod_{t=2:T} \left\{P(F_{t-1}^{1:H}|L_t^{1:H}) P(L_t^{1:H}|L_{t-1}^{1:H}, F_{t-1}^{1:H}) P(L_{t-1}^{1:H}, F_{t-1}^{1:H})\right\}}_{\text{Conditional PDF at the successive time frames}}$$

6.4.4 Inference in H-SLDSs

The exact inference in H-SLDSs is infeasible due to the the intrinsic intractability of inference for SLDSs where the number of Gaussian mixtures to represent the uncertainty for the continuous states X increases exponentially over time [45]. In terms of the hierarchical data interpretation tasks, the inference step aims to compute the posterior distribution on the entire discrete states at all the hierarchies $P(L_{1:T}^{1:H}, F_{1:T}^{1:H}|Z)$ as accurately as possible. An example method used in this work include an approximate Viterbi method [69] which approximates the target distribution $P(L_{1:T}^{1:H}, F_{1:T}^{1:H}|Z)$ with a single most-likely label sequence $P(L_{1:T}^{1:H}, F_{1:T}^{1:H}|Z) \approx \delta(\hat{L}_{1:T}^{1:H}, \hat{F}_{1:T}^{1:H})$ and a variational approximation method [69] which computes the probabilistic posterior distribution under the lower-bound optimization principle.

In this work, we re-used the available inference methods for standard SLDSs for H-SLDSs. This is possible because we can convert an H-SLDS model into an equivalent flat SLDS model by collapsing the hierarchy and create meta-variables by merging the discrete states and the finish variables. Eventually, the HMC in the original H-SLDS model collapses to a flat Markov chain, and we obtain an equivalent SLDS model. The resulting SLDS model has a discrete state space with a size of order $O(h \times \prod_{h=0}^{H} |L^{(h)}|)$ where the first term and the second product term correspond to the size of the configurations for the finish variables and the discrete states respectively where $|L^{(h)}|$ denotes the cardinality for the h-th layer. While such exponential increase in state size may look infeasible, the resulting collapsed Markov chain is mostly sparse since every parent state invokes only a small subset of the available sub-states. Hence, if we use a sparse matrix representations to encode the resulting Markov chain, the size of the model is often found to be well-tractable within manageable bounds.

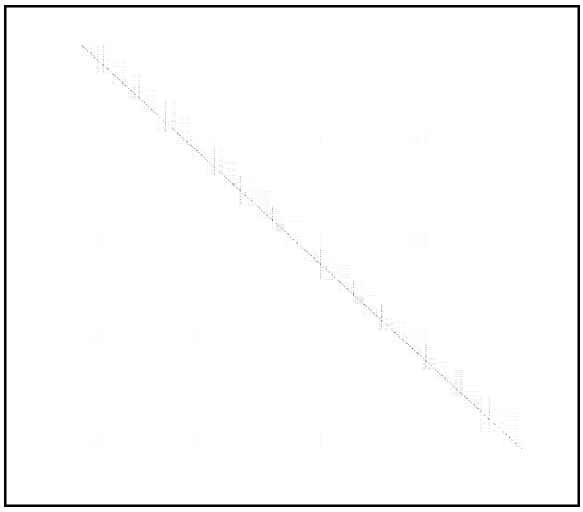


Figure 24: An illustration of a flattened Markov chain with total 9,450 states obtained by flattening the hierarchical Markov chain of H-SLDSs during the conversion process. Darker color represent higher probability areas where the rows and the columns correspond to previous and next states respectively. It can be observed that the resulting transition model is *sparse*.

This model conversion approach was also successfully used for S-SLDSs, as presented in Chapter. 3.

The time-complexity of the above approximate inference methods are strictly linear w.r.t. the time and roughly linear w.r.t. the meta-state size. The roughly linear (rather than quadratic) computational overload w.r.t. the state size is due to the fact that the state transitions are sparse and mostly bounded by a modest constant that arises from the structured topology within HMC, i.e, the maximum number of transitions for an HMC state is bounded in practice. Once the posterior on the discrete states are computed in the equivalent SLDS representation, the posterior can be marginalized and the targeted posterior on multiple layers can be obtained. To illustrate the sparseness of the equivalent flat Markov chain, we converted a 3-layer H-SLDS model into an equivalent SLDS model where the resulting flattened Markov chain possessed 9,450 states. To illustrate the sparseness of the transitions, the obtained flat transition matrix is depicted in Fig. 24. It can be observed that the transition model is extremely sparse. Using the sparse matrix routines which effectively discards the zero probability transitions, the inference in the large H-SLDS models could be conducted in reasonable time in practice with a gain in the speed to the order of three.

It is worth noting that the worst case scenario of exponential increase in the size of the meta states is limited in practice when the hierarchical models are applied to appropriate problems with sparse grammar-like structures. The grammar-like structure between the states in the parent layer and the child layer is sparse in most cases, and avoids the exponential increase in the flattened states. As an example, we can assume that there are two child states 'a' & 'b', which are initiated by the parent states 'A' & 'B' respectively. Now the flattened states are 'Aa' and 'Bb' (constant), instead of 'Aa','Ab','Ba','Bb' (exponential increase). As can be seen from this example, the exponential increase is well avoided, unless naively done. The provided exponential increase in the number of states is indeed only the worst case where all the parent layers are densely connected to child layer. In fact, we can argue that densely connected models do not exploit the advantage of the sparse essence of the hierarchical models.

Furthermore, the use of flattening procedure to produce an equivalent SLDS model bears a benefit that it excludes additional levels of approximation in the inference results. Such methodology will help us to carefully examine the power of the model as close to the ground truth as possible.

6.4.5 Learning in H-SLDSs

The learning problem for H-SLDSs consists of two sub-problems : (1) the parameter learning problem given a fixed model structure and (2) the structure learning problem.

In terms of the parameter learning problem, an EM-based approach can be used to improve the model parameters under the maximum-likelihood principle. Based on the success of the EM-based parameter learning for H-HHMs [29, 56, 43, 27] and SLDSs [37, 91, 64, 63], the EM-based learning method has been used as the primary learning machinery for our work on H-SLDSs.

In our work of applying H-SLDSs to the problems of automatic human exercise annotation presented in Chapter 7, we used a hybrid learning approach where we provided a set of labeled training data for the upper layers of the hierarchy and the parameters for the lower layers are learned in an unsupervised manner.

The structure learning problem for H-SLDSs has been only partially addressed within limited scope where we take a bottom-up learning approach and used empirical measures to determine the size of the LDS vocabularies. As mentioned in Section 6.2, different number of LDSs were learned and the maximum size where every cluster represents more than certain amount of data was chosen as the underlying structure. For the upper layers and LR-SLDSs, we used available domain knowledge or hand-turning to grasp reasonable model structures. For the data where neither the existence nor the structure of the HMC is unclear, we may have to rely on the unsupervised structure learning work. However, we plan to address the unsupervised structure learning problem for the deep hierarchies as part of our future work.

6.5 Discussion and Related work

6.5.1 Direction for more Scalable Inference method

It would be worth noting that a computationally *less* demanding inference algorithm for the presented H-SLDS model may be derived under certain approximation assumptions, e.g., a structured variational approximation [39] may be used, as has been done for another extension of SLDSs [37]. Howard and Jebera [37] provided an iterative variational inference method for the 'dynamical system tree' model where the overall posterior distribution over the hidden variables is approximated as the product of probability distribution over individual Markov chains at multiple layers. By adopting such strategy, the demanding computation on the whole hierarchical Markov chain divides into a set of smaller Markov chain computation at every layer. While such approach would introduce more approximation to the exact joint posterior distribution, it would be desirable to bear such disadvantage and gain speed and reduced memory requirements to apply H-SLDSs to overly large problems where there would be far more than a million state configurations. Although the structured variational inference algorithm for H-SLDSs is not presented in this dissertation, we plan to address the problem for H-SLDSs and H-HHMs [56] in general in the future.

6.5.2 Learning in H-SLDSs

It would be worth noting that one would be able to obtain the optimal learning results for H-SLDSs when the amount of unsupervised learning is reduced as much as possible. The

argument is that we often have semantic concepts related to the upper layers of the hierarchical model, and we should exploit such known relationships between semantic concepts by providing labeled data as much as possible. In detail, it would be best to provide labeled datasets for the upper layers to facilitate the learning procedure in a supervised setting. In parallel, the low-level primitive patterns which underlie the semantic concepts can be learned from the data in an unsupervised manner within EM framework, as it has been done to learn LR-SLDSs in Section 6.3. In particular, since the upper layers correspond to coarser time scales and possess less number of states and associated parameters, there is a high possibility that amount of training data to learn meaningful and generalizable parameters is within practical range. Moreover, every time we reduce the number of layers whose parameters are to be learned in an unsupervised manner, the size of the meta state size decreases exponentially, and so does the amount of training data. While earlier work has shown certain empirical success of hierarchy learning in a full unsupervised setting [43, 37], the problem still remains as finding a needle in a huge haystack. Furthermore, we would need substantial progress in the methodologies to test the generalization ability of the learned models across different datasets and to provide debugging ability to demonstrate the learned concepts in a human-understandable forms.

6.5.3 Representational Power of H-SLDSs

The probabilistic automata encoded by the H-SLDS model corresponds to probabilistic hierarchical regular grammar whose representational power is described best by a form of non-deterministic push-down automata with bounded-sized queues. Specifically, the finish variables act as the elements in the context queues but the size of the queue in the H-SLDS model is bounded because the hierarchy is of fixed height.

A temporal model representation which is more powerful than the hierarchical Markov chain in terms of the representational scope is a stochastic context-free grammar (SCFG), which is also called a probabilistic context-free grammar (PCFG). SCFG is a context-free grammar in which each production is augmented with a probability. The probability of a derivation (parse) is then the product of the probabilities of the productions used in that

derivation; thus some derivations are more consistent with the stochastic grammar than others. SCFGs extend context-free grammars in the same way that hidden Markov models extend regular grammars. SCFGs have application in areas as diverse as from natural language processing to the study of RNA molecules. For example, previous work [38, 54] used SCFG to interpret the high-level behaviors from video measurements, such as car parking, recognition of rectangular shape drawings, and card plays.

While the class of SCFGs provide superior representational power, the major challenge for the labeling problem is that the worst case asymptotic time complexity of the inference algorithm such as Cocke-Younger-Kasami (CYK) algorithm (alternatively called CKY) is $O(T^3)$ w.r.t. the length of the sequences. In particular, the worst case time complexity holds when the interpretation (parsing) of data is ambiguous, i.e., there can be multiple interpretations available for a sequence, which is often the case when we apply the model to a long sequence and the model is stochastic. When a context free grammar is unambiguous, the time complexity reduces to $O(T^2)$, and the complexity eventually becomes linear $O(T)$ when all the rules are left-to-right, which is essentially a regular grammar, which is identical to the inference complexity of the presented H-SLDS model.

In practice, a large number of hierarchical phenomenon can be effectively and sufficiently encoded using hierarchical regular grammars, i.e., hierarchical Markov chains. In particular, the semantic concepts that we are interested in practice have mostly bounded depth of grammar composition hierarchy, probably due to the limit of human reasoning capabilities. Indeed, the examples in many works on SCFGs for probabilistic interpretation of data can be successfully modeled using hierarchical Markov chains. A major representational benefit of using SCFGs is that it allows us to capture the recursive structures. For example, a context free grammar can represent recursively nested productions such as $A \rightarrow aAb \rightarrow aaAbb$ while regular grammars can not. However, such heavily recursive nesting patterns exist very rarely (with the exception in biological systems such as DNAs). When the recursiveness is bounded, most of such nested structures can be represented as left-to-right structures which are easily encoded by regular grammars. Hence, the use of simpler hierarchical Markov chain is often sufficient for many practical problems and we should avoid the use of more complex

SCFGs unless there are pertinent need to pursue such direction, in particular when we are dealing with data captured over a long period of time.

Chapter VII

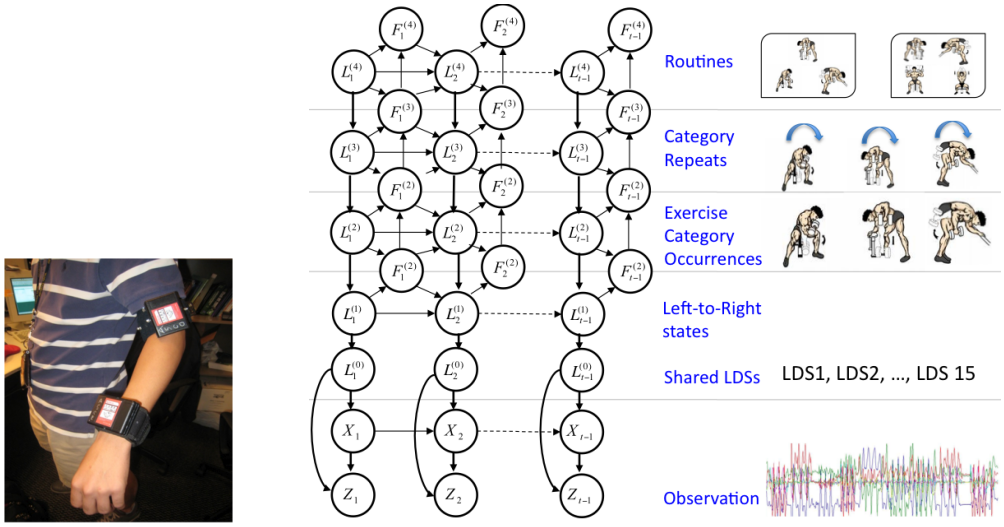
AUTOMATED ANNOTATION OF EXERCISES USING H-SLDSS

The application to which we apply H-SLDSs is the automated human exercise annotation. This task is important for the health monitoring industries where the system would automatically generate a report on the amount and the categories of exercises the subjects conducted, which would eventually save the substantial cost of appointing professional personnels simply to monitor the exercise progress frequently. Moreover, the subject individuals would be able to keep track of their own progress over time, keep the history of their exercises, and review the stored information in the future.

7.1 Two Dumbbell Exercise Datasets and H-SLDSs

Two dumbbell exercise datasets were used to test the practical usefulness of the developed H-SLDS model. The first dataset was collected in [53] where an XSens MT9 inertial motion sensor was attached to the subject's wrist by fitting it into a pouch sewn to the back of a thin glove. The original data is sub-sampled at 12.5Hz where three axis accelerometer and gyroscope readings are recorded to collect the entire 31 sequences. The second dataset was collected by the author using two wireless bluetooth accelerometer sensors attached to the upper and lower arms respectively, where 5 different subjects conducted 6 sequences each, following the choreographies designed by the author. A snapshot of the two bluetooth sensors attached on an arm of a subject is shown in Fig. 25 (a). Both datasets are six dimensional where a single accelerometer sensor used for the first dataset recorded three dimensional directions and acceleration in each dimension, which sums to six dimensional data. On the other hand, the two bluetooth sensors used to record the second dataset reported only the acceleration in three dimensions, which sums to six dimensional data again.

Once the datasets were collected, they were manually labeled at all hierarchies down to every occurrence of an exercise to provide ground truth (GT) labels against which the accuracy of the developed H-SLDS framework is measured. In particular, we have recorded



(a) (b)

Figure 25: Use of H-SLDSs for the human exercises. (a) A snapshot of the two bluetooth sensors attached on an arm of each subject. (b) A 4-layer H-SLDS model for dumbbell exercises. The hierarchies correspond to the exercise routines, exercise repeats, each exercise occurrence, and LR-SLDS states, from top to the bottom layer respectively.

the exercise of the subjects using video-cameras which are synched with motion signals from the wearable sensors, and these video records were used to aid to mark the ground-truth labels of the collected sequences.

The hierarchic structure in both datasets are similar where the corresponding DBN for the H-SLDS model is shown in Fig. 25(b). In both datasets, the subjects conducted six different dumbbell exercises (plus one unknown) illustrated in Fig. 19. Accordingly, there are seven states in the second layer (from bottom) of the DBN where each state corresponds to an occurrence of an exercise. The third layer of the DBN captures the repeatedness of the conducted exercises : each subject was asked to repeat an exercise three times successively whenever they conduct distinct exercises. Hence, there are seven repeat states in the third layer, identical to the second layer. However, the states in the third layer do not switch even if an identical exercise repeats successively while the second layer would switch. Finally, at the top level, there are 'routines' which are sequences of three distinct exercise categories.

For example, a routine would be a sequence of three flat curls, three shoulder extensions, and three triceps. The two datasets contain four and six routines respectively where the top layer of the DBN comprises of these routine states. Every sequence was choreographed to

Table 5: The characteristics of the two exercise datasets.

	Data 1	Data 2
No. of On-body Sensors (Type)	1 (Wired)	2 (Bluetooth)
Noise level	Low	High
Sampling Frequency	12.5 Hz	30 hz
No. of Human Subjects	1	5
No. of Choreographies	4	6
Total No. of Sequences	31	30
No. of Routines	4	4
No. of Exercise Categories	6 (+1 unknown)	6 (+1 unknown)
Leave-one-out Criteria	Every sequence	Every subject

contain exactly three routines. By choreography, we mean the pre-designed exercise scripts which the participating subjects were asked to follow. In summary, the characteristics of the two datasets are tabulated in Table 5. Note that, in the first dataset, a single subject conducted an identical choreography multiple times (average 7 times). On the other hand, five subjects conducted every choreography exactly once in the second dataset.

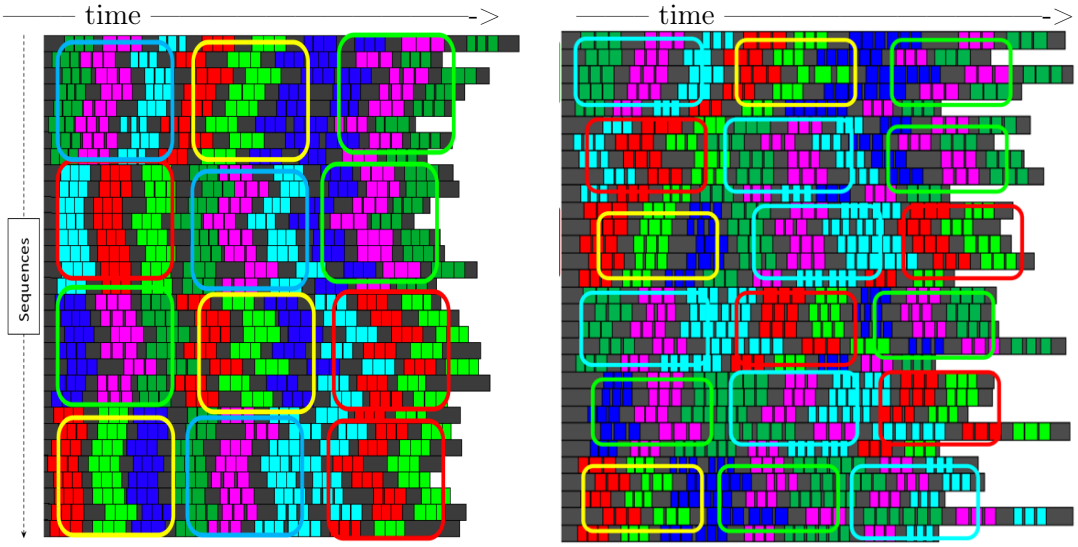
More in detail, the first dataset consists of the following four choreographies :

$$\begin{aligned}
\text{Choreogrpahy 1} &\triangleq [\text{ Routine 1; Routine3; Routine 4}] \\
\text{Choreogrpahy 2} &\triangleq [\text{ Routine 2; Routine1; Routine 4}] \\
\text{Choreogrpahy 3} &\triangleq [\text{ Routine 4; Routine3; Routine 2}] \\
\text{Choreogrpahy 4} &\triangleq [\text{ Routine 3; Routine1; Routine 2}]
\end{aligned}$$

The second dataset comprises of the following six top-level choreographies :

$$\begin{aligned}
\text{Choreogrpahy 1} &\triangleq [\text{ Routine 1; Routine 3; Routine 4}] \\
\text{Choreogrpahy 2} &\triangleq [\text{ Routine 2; Routine 1; Routine 4}] \\
\text{Choreogrpahy 3} &\triangleq [\text{ Routine 3; Routine 1; Routine 2}] \\
\text{Choreogrpahy 4} &\triangleq [\text{ Routine 1; Routine 2; Routine 4}] \\
\text{Choreogrpahy 5} &\triangleq [\text{ Routine 4; Routine 1; Routine 2}] \\
\text{Choreogrpahy 6} &\triangleq [\text{ Routine 3; Routine 4; Routine 1}]
\end{aligned}$$

The four routines shared in both dataset are as follows where the patterns in parenthesis $(\cdot)$ indicate that those patterns may occur by chance :



(a) Dataset 1

(b) Dataset 2

(c) The color map of seven different exercise categories.

Figure 26: The color-coded visualization of the datasets. Every row represents distinct sequences. (a) Dataset 1. (b) Dataset 2. (c) The color map of seven different exercises. The exercises correspond to flat curl, shoulder extension, back, twist curl, shoulder press, tricep, and unknown, from left to right. It can be observed that there are four and six different choreographies for the first and the second dataset respectively.

$$\text{Routine 1} \triangleq [(\text{Repeat 7});\ \text{Repeat 1};(\text{Repeat 7});\ \text{Repeat 2};(\text{Repeat 7});\ \text{Repeat 3};(\text{Repeat 7})]$$

$$\text{Routine 2} \triangleq [(\text{Repeat 7});\ \text{Repeat 3};(\text{Repeat 7});\ \text{Repeat 4};(\text{Repeat 7});\ \text{Repeat 5};(\text{Repeat 7})]$$

$$\text{Routine 3} \triangleq [(\text{Repeat 7});\ \text{Repeat 4};(\text{Repeat 7});\ \text{Repeat 5};(\text{Repeat 7});\ \text{Repeat 6};(\text{Repeat 7})]$$

$$\text{Routine 4} \triangleq [(\text{Repeat 7});\ \text{Repeat 6};(\text{Repeat 7});\ \text{Repeat 2};(\text{Repeat 7});\ \text{Repeat 1};(\text{Repeat 7})]$$

Finally, each of the repeat patterns shared between two datasets correspond to the successive occurrence of exercise categories where there are seven repeat patterns that correspond to the six exercise categories and additional unknown pattern. The patterns from Repeat 1 to Repeat 7 correspond to the repeated occurrences of flat curls, shoulder extensions, backs, twist curls, shoulder presses, triceps, and unknowns.

To clarify the hierarchic structure within data further, two color-codings of the datasets are shown in Fig. 26 where every row represents a sequence and every colored rectangle corresponds to an occurrence of a particular exercise. The larger rounded rectangles overlaid across multiple sequences roughly identify the four different routines colored in cyan, red, yellow, and green respectively. Additionally, Note that the illustrated labels are manually obtained ground truth information.

7.1.1 Learning H-SLDSs for the Exercise Datasets

The H-SLDS model used to automatically annotate the dumbbell exercise datasets is shown in Fig. 25. The parameters of the hierarchical Markov chain of the H-SLDS model are initialized based on the domain knowledge described in the Section 7.1. In terms of shared LDS vocabulary, total of 13 and 15 LDSs are learned from the first and second data respectively, based on the mixture modeling approach described in Section. 6.2. Then, 7 different LR-SLDSs for each category are learned on top of the LDS vocabulary, based on the approach described in Section. 6.3. Now that all the parameters are available, H-SLDSs are formed for both datasets and converted to equivalent SLDSs.

7.1.2 Inference in H-SLDSs

Once the H-SLDS models are constructed, the models are applied to automatically annotate novel test sequences using two inference methods : (1) a variational approximation (VA) method [69, 60] generates probabilistic posterior distributions, and (2) an approximate Viterbi inference method (VI) generates a single most-likely label sequence. The inference task, i.e., data annotation, is conducted using the equivalent SLDS models and the results are marginalized through post-processing to be converted back to hierarchical representations. The resulting hierarchical annotations contain the interpretation of data at multiple layers simultaneously. We would show the results through visualization in Section 7.3.

7.2 Experimental Results

During the training phase described in Sec. 7.1.1, we adopted the leave-one-out approach (LOO) - we excluded a part of the data from the training phase and tested the learned

model on the excluded dataset. Such experiments are repeated for all possible divisions of the datasets. For the first dataset, we conducted LOO experiments at the sequence level : every sequence is excluded in turn within the LOO testing framework, resulting in total of 31 experiments. For the second dataset, the LOO experiments are conducted at the subject level : the data belonging to a particular subject was excluded and tested after the parameters are learned from the remaining data, resulting in total of 5 experiments. To understand the impact of the hierarchy, H-SLDS models are built for the second dataset without 'Repeat' and 'Routine' layers, and experiments were conducted to label the sequences in the identical LOO setting.

In the following sections, both qualitative and quantitative results are presented in Sec. 7.3 and Sec. 7.4 respectively.

7.3 Qualitative Results

The experimental results demonstrate that human exercise datasets can be successfully interpreted across multiple time resolutions. Among the experimental results, four representative satisfactory results by H-SLDSs from each dataset are shown in Figures 27 and 28 respectively. For each result, the smoothed posterior by variational approximation (VA), the most-likely annotation by approximate Viterbi (VI), and the ground truth (GT) are color-coded for every layer where the layers correspond to the routines, repeats, and individual occurrence from top to the bottom. The x-axis represents time-flow and the color is the label at that corresponding time frame. The vertical black bars in the VI results shown in the middle color strips for all the layers denote the switchings detected by H-SLDSs. The color maps for the 'Repeat' and 'Category' levels refer to Fig. 26(c). Note that a different color map is used to visualize the four and six 'Routines' at the top level for each dataset respectively.

For the first dataset, the inference results for all the 31 sequences were fairly good and comparable to the representative results shown in Fig. 27. It can be observed that the smoothing results (VA) show plausible soft labeling results which closely match the ground truth. In particular, there were no particularly unsuccessful result for the first

dataset. Although such overall successful results are satisfying, it is worth noting that the first dataset was entirely collected from a single subject where even the applied LOO setting does not prevent the model from learning fairly large amount of information about the subject through the separate training datasets. Hence, the successful results on the first dataset demonstrates that H-SLDSs may be used successfully per user-basis once enough data is provided. More insight about the generalization power of H-SLDSs can be obtained by analyzing the results from the second dataset.

For the second dataset, we obtained successful results in general where four representative examples are shown in Fig. 28. The second dataset is more challenging than the first dataset since (1) the experiments were conducted through a subject-level LOO setting where the data from the test subject was never used in the training stage, and (2) the sensor measurements consist of only acceleration measurements obtained from two more noisy wireless sensors and the very distinctive orientation measurements in the first dataset were missing. It is interesting to note that the smoothing results for the second dataset in Fig. 28 are much sharper than the ones shown for the first dataset in Fig. 27. While it is hard to exactly understand why the resulting posterior distributions are sharper for the second dataset, we conjecture that the parameters of H-SLDSs learned for the second dataset are less fine-tuned for the test datasets, and tend to produce more over-confident labeling results.

To examine the usefulness of the high-level information, we have built a shorter H-SLDSs (with less hierarchy) which only models up to the category occurrence layer. Exactly the same category models are used, but the Markov switching behavior between the categories were assumed to be uniform, including the unknown category. The labeling results obtained by this simpler model is shown in Fig. 29 where the sequences are identical to the ones shown in Fig. 28. It can be observed that the amount of additional error due to the absence of higher-level information such as repeating patterns and high-level routines are substantial. In particular, the unknown categories are almost never detected and there is high ratio of substitution errors between back (cyan), flat-curl (red), and twist-curl (green) exercises.

Additionally, there were several unsatisfactory experimental results for the second dataset. A set of unsatisfactory results for the second datasets are shown in Fig. 30 where we can

Table 6: Gender and Heights of the subjects involved in the data collection for Dataset 2. The bottom row shows gender where M and F denotes male and female respectively.

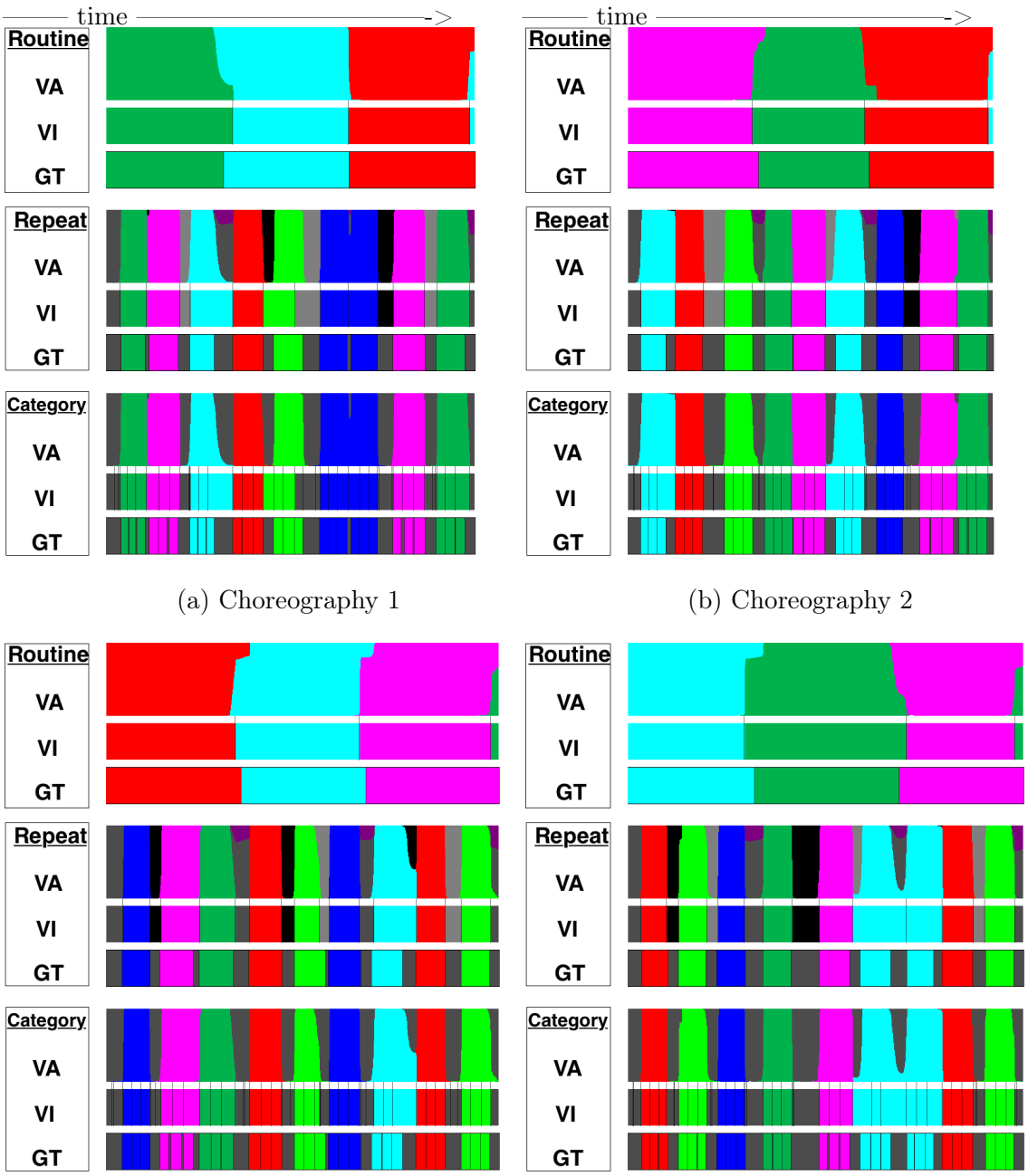
Subjects	1	2	3	4	5
Height (cm)	190	160	175	179	173
Gender (M / F)	M	F	M	M	M

observe different types of errors such as insertion, substitution and shift errors. In terms of insertion and substitution errors, these errors were introduced mostly when the following exercise categories were confused : flat curl (green), twist curl (red), and back (cyan). The sources of such confusion can be further analyzed by revisiting the interpreted LDS labels for each category shown in Fig. 22 - it can be observed that the interpreted LDS switching patterns for these confused categories show high similarity. In particular, the only difference in the two exercises, flat curl and twist curl, is the slight wrist twist at the top of the lifting motion for the twist curl exercise while we keep the wrists unmoved (flat) for the flat curl exercise. In terms of the confusion w.r.t. the back exercise, it can be guessed that the sensor measurements may not be substantially different from the curl-type exercises given that there were no orientation readings. Furthermore, Table 6 shows that the subjects who participated in the data collection for the second dataset presented very different physical characteristics where the height of the subjects ranged from 160 cm to 190 cm. And the gender of the subjects were mixed where the second subject was a female and the others were male. The substantial difference in physical characteristics of the subjects may have caused H-SLDSs to generalize less given the sparse dataset. It is worth noting that the poorest results were obtained from the data obtained from the second subject who was the only female and whose height was most apart from the median of the heights of the participants.

7.4 Quantitative Results

The experimental results are quantitatively analyzed by computing the matching accuracy which is the ratio of the correctly inferred VI labels w.r.t. the ground truth (GT) across all the time frames and layers, where the accuracy ranges from zero (total failure) to one (perfect match). Table 7 shows the detailed accuracy results for both datasets.

The average accuracy was 85% for the first dataset and 76% for the second dataset.



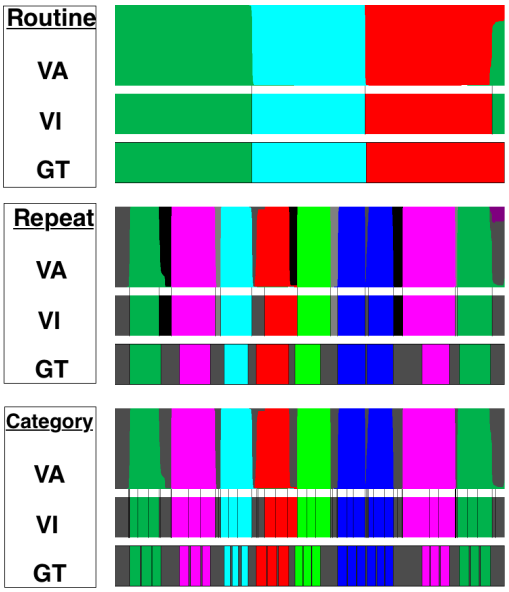
(a) Choreography 1

(b) Choreography 2

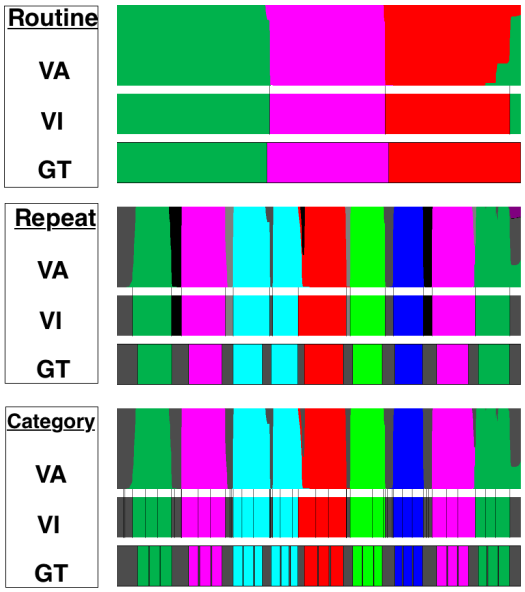
(c) Choreography 3

(d) Choreography 4

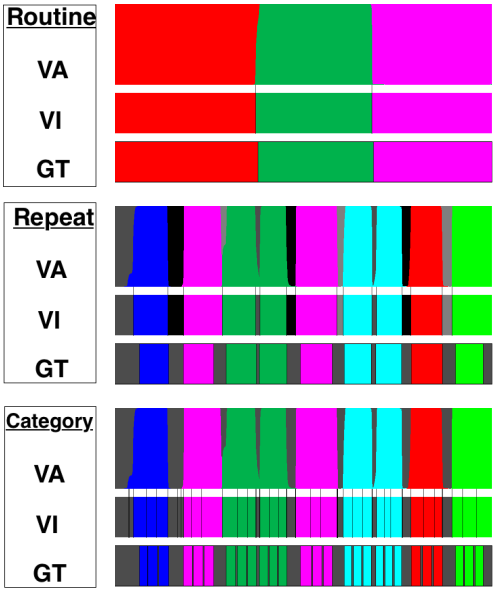
Figure 27: ***Satisfactory hierarchical labeling results from Dataset 1.*** Representative examples for each of the four choreographies are shown. For each result, the smoothed posterior by variational approximation (VA), the most-likely annotation by approximate Viterbi (VI), and ground truth (GT) are color-coded for every layer where the layers correspond to the routines, repeats, and individual occurrence from top to the bottom. The color maps for the 'Repeat' and 'Category' levels refer to Fig. 26(c). Note that a different color map is used to visualize the four 'Routines' at the top level.



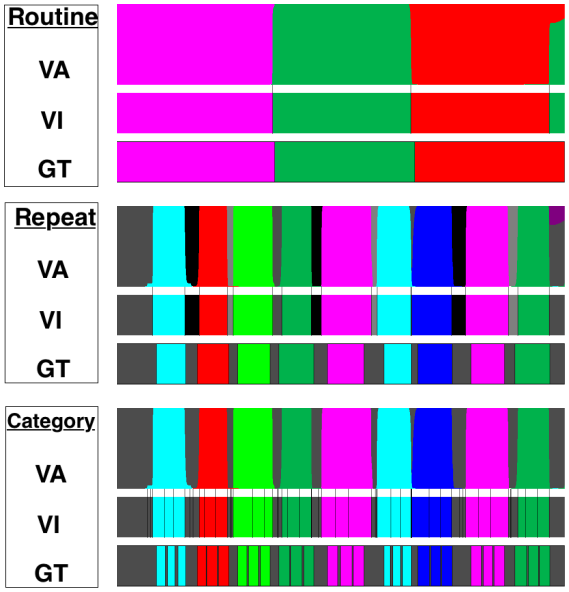
(a) Sub. 1, Ch. 1



(b) Sub.1, Ch. 4

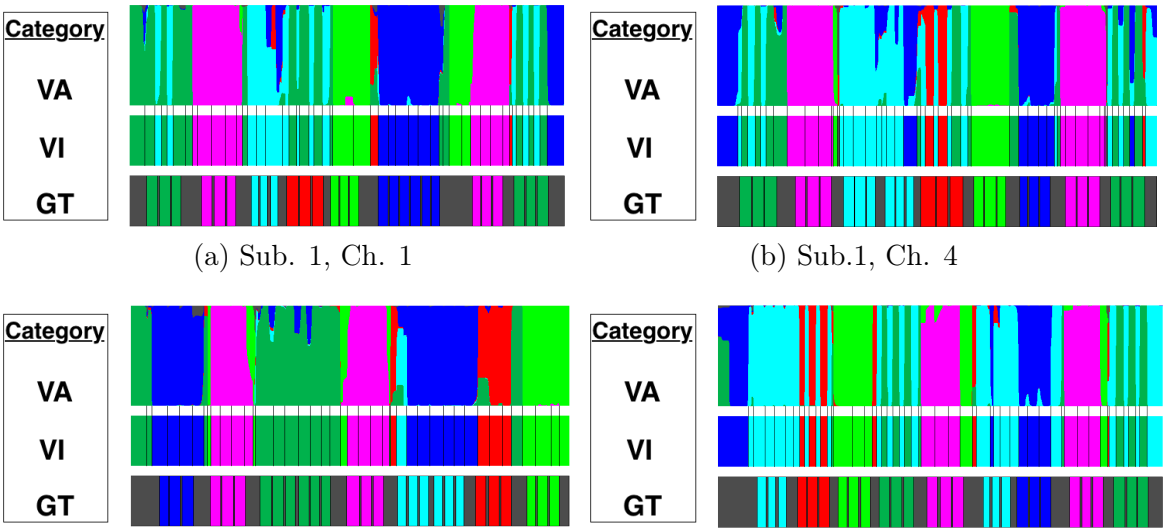


(c) Sub. 3, Cho. 5



(d) Sub. 4, Cho. 2

Figure 28: ***Satisfactory hierarchical labeling results from Dataset 2.*** Representative examples from different subjects and choreographies are shown. For each result, the smoothed posterior by variational approximation (VA), the most-likely annotation by approximate Viterbi (VI), and ground truth (GT) are color-coded for every layer where the layers correspond to the routines, repeats, and individual occurrence from top to the bottom. The color maps for the 'Repeat' and 'Category' levels refer to Fig. 26(c). Note that a different color map is used to visualize the six 'Routines' at the top level. Beneath each result, the subject and the corresponding choreography are shown.



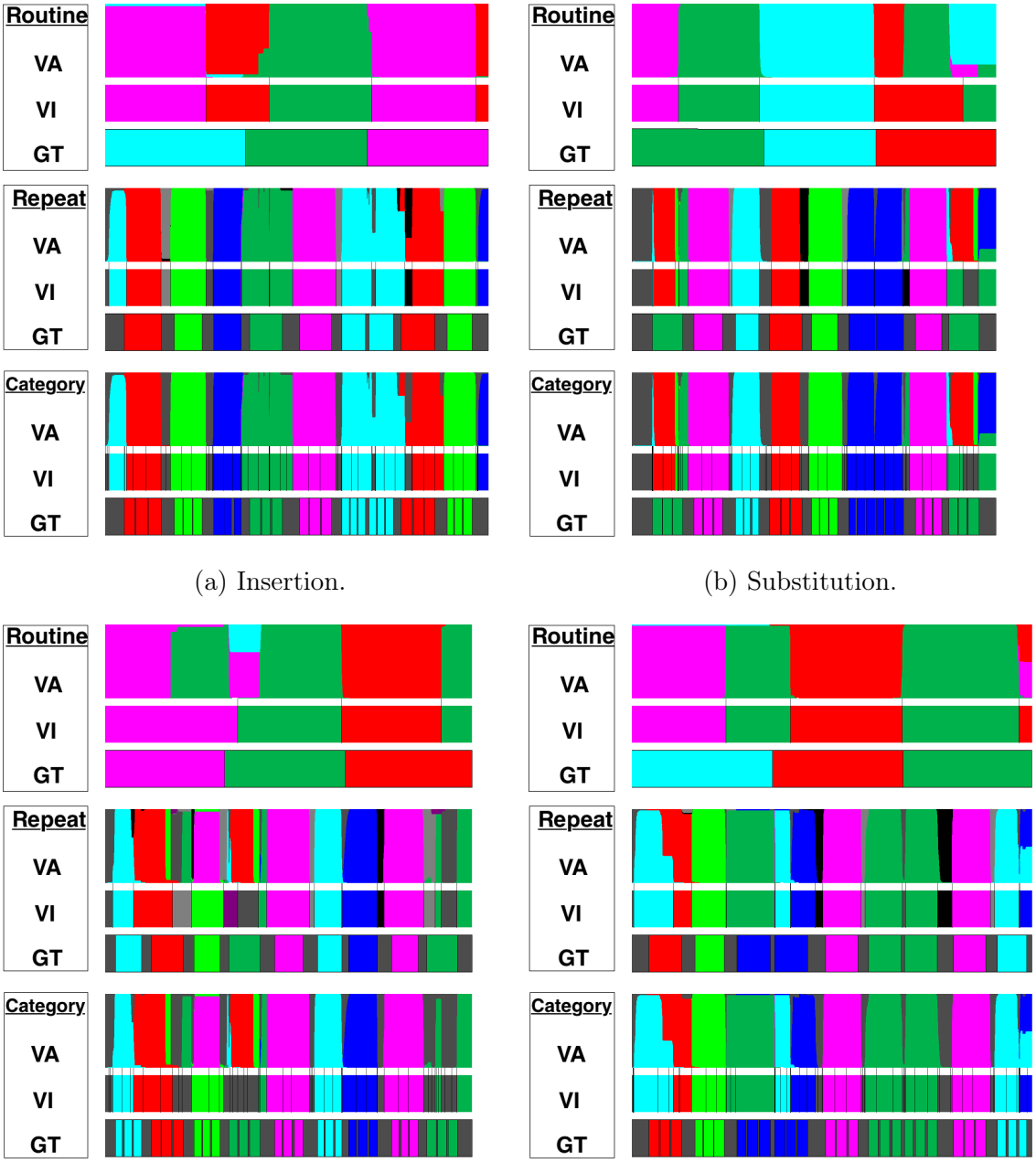
(a) Sub. 1, Ch. 1 (b) Sub.1, Ch. 4

(c) Sub. 3, Cho. 5 (d) Sub. 4, Cho. 2

Figure 29: ***Labeling results from Dataset 2 with shorter H-SLDSs.*** The labeling results obtained for the sequences shown in Fig. 28. A shorter H-SLDS model with hierarchy only up to category layer was built, and used to label the data. For each result, the smoothed posterior by variational approximation (VA), the most-likely annotation by approximate Viterbi (VI), and ground truth (GT) are color-coded for the category layer. It can be seen that the amount of additional error due to the absence of hierarchy is substantial.

The accuracy for the both datasets are fairly good. Furthermore, we can highly esteem even the lower accuracy for the second dataset since the dataset poses more challenges as mentioned in Section 7.3. Moreover, a large portion of the errors is due to the slight boundary misalignment between the VI and GT labels, as it can be observed in Fig. 28. In contrast, the overall accuracy of the shorter H-SLDS model which encodes only up to the category layer was 49% for the second dataset.

In Table 7 (a), the results for the first dataset are organized in choreography-wise order. The double-line borders between the rows indicate the boundaries between the subjects that belong to different choreographies. However, distinctive difference between accuracy for different choreographies can not be observed. For the experimental results for the second dataset shown in Table 7 (b), the results are organized in subject-wise order. The top six results belong to the first subject, and the bottom six results belong to the fifth subject. The double-line borders between the rows indicate the boundaries between the subsets which belong to different subjects. The poorest results are highlighted in bold fonts. It can be



(a) Insertion.

(b) Substitution.

(c) Shift.

(d) Substitution & Insertion.

Figure 30: ***Unsatisfactory hierarchical labeling results from Dataset 2.*** Qualitative errors are identified by the difference between the Viterbi labels and ground truth labels. Insertion, substitution, and shift errors are most common error types. For each result, the smoothed posterior by variational approximation (VA), the most-likely annotation by approximate Viterbi (VI), and ground truth (GT) are color-coded for every layer where the layers correspond to the routines, repeats, and individual occurrence from top to the bottom. The color maps for the 'Repeat' and 'Category' levels refer to Fig. 26(c). Note that a different color map is used to visualize the six 'Routines' at the top level. Beneath each result, the types of errors are shown : insertion, substitution, and shift errors.

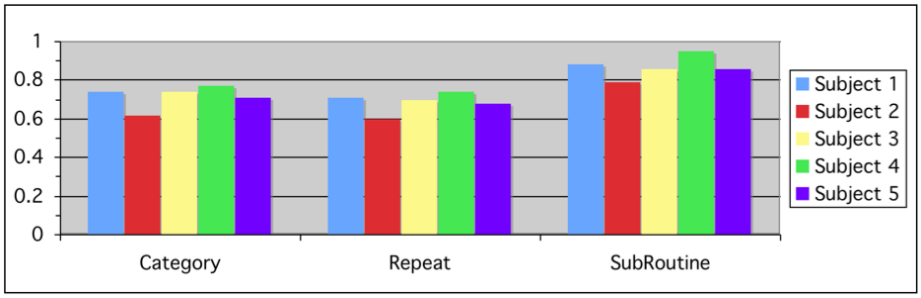


Figure 31: Subject-wise accuracy results for Dataset 2. It can be seen that the test results on the data from the second subject shows lowest accuracy across all the layers.

observed that most of the unsatisfactory results are obtained from the data collected from the second subject. A graph in Fig. 31 highlights that the accuracy for the second subject who is the only female and whose height is the most different from the median heights of all the subjects is lowest, which suggests that we would need larger and denser amount of training dataset to achieve superior generalization power using H-SLDSs.

7.5 Conclusion and Related Work

We presented our experimental results on the real-world human exercise datasets, where the two exercise datasets were modeled using H-SLDSs. Both the qualitative and quantitative results demonstrate that the presented H-SLDS model can infer the labels across multiple layers fairly accurately. The two datasets exhibit different types of sensor measurements and have been collected using different physical sensors, which suggests that H-SLDSs can be useful in variety of settings and sensor modalities. In particular, the experimental results on the second dataset demonstrate that H-SLDSs provide a working framework which can generalize to label data from unseen subjects.

We believe that the experimental results presented in this chapter is one of the most thorough results to demonstrate the practical usefulness of the hierarchical temporal models, in comparison to the previous work which dealt with toy examples [56, 91] or problems without ground truth [37, 43]. A related work [17] on using H-HHMs for vision-based activity and plan recognition also presented successful results on fairly large problems in terms of hierarchy complexity and the number of concept classes, however, the length of their sequences were often substantially shorter than the results reported in this chapter

Table 7: Accuracy results for the two datasets. (1) **Results for Dataset 1**. The results are organized in choreography-wise order. The double-line borders between the rows indicate the boundaries between the subjects that belong to different choreographies. Distinctive difference between accuracy for different choreographies can not be observed. (b) **Results for Dataset 2**. The results are organized in subject-wise order. The top six results belong to the first subject, and the bottom six results belong to the fifth subject. The double-line borders between the rows indicate the boundaries between the subsets which belong to different subjects. The poorest results are highlighted in bold fonts. It can be observed that most of the unsatisfactory results are obtained from the data collected from the second subject. A bar graph which shows subject-wise accuracy across different hierarchies are shown in Fig. 31.

Sequence	Category	Repeat	Routine	Average
1	0.84	0.73	0.92	0.83
2	0.89	0.77	0.93	0.86
3	0.93	0.71	0.96	0.87
4	0.81	0.68	0.94	0.81
5	0.91	0.77	0.94	0.87
6	0.75	0.73	0.96	0.81
7	0.81	0.73	0.94	0.83
8	0.80	0.74	0.92	0.82
9	0.84	0.78	0.95	0.86
10	0.82	0.77	0.94	0.84
11	0.89	0.80	0.96	0.88
12	0.82	0.80	0.95	0.85
13	0.87	0.77	0.94	0.86
14	0.89	0.82	0.94	0.88
15	0.87	0.76	0.93	0.85
16	0.86	0.82	0.95	0.88
17	0.87	0.77	0.95	0.86
18	0.88	0.80	0.95	0.88
19	0.90	0.78	0.94	0.87
20	0.87	0.77	0.94	0.86
21	0.90	0.75	0.93	0.86
22	0.85	0.78	0.92	0.85
23	0.78	0.74	0.94	0.82
24	0.81	0.75	0.91	0.82
25	0.87	0.74	0.93	0.85
26	0.88	0.75	0.92	0.85
27	0.85	0.76	0.96	0.86
28	0.82	0.76	0.91	0.83
29	0.82	0.73	0.93	0.83
30	0.78	0.69	0.96	0.81
31	0.81	0.81	0.95	0.86
Average	0.85	0.76	0.94	0.85

(a) Dataset 1

Sequence	Category	Repeat	Routine	Average
1	0.73	0.70	0.96	0.80
2	0.77	0.73	0.90	0.80
3	**0.66**	**0.67**	**0.53**	**0.62**
4	0.81	0.78	0.96	0.85
5	0.71	0.68	0.96	0.78
6	0.75	0.70	0.96	0.80
7	0.66	0.64	0.76	0.69
8	**0.54**	**0.51**	**0.87**	**0.64**
9	0.71	0.65	0.92	0.76
10	**0.61**	**0.60**	**0.76**	**0.66**
11	0.66	0.64	0.81	0.70
12	**0.57**	**0.59**	**0.62**	**0.59**
13	0.79	0.76	0.97	0.84
14	0.64	0.61	0.81	0.69
15	0.76	0.75	0.56	0.69
16	0.74	0.69	0.91	0.78
17	0.80	0.75	0.99	0.85
18	0.71	0.64	0.89	0.75
19	0.80	0.75	0.95	0.84
20	0.79	0.75	0.95	0.83
21	0.69	0.70	0.96	0.78
22	0.76	0.73	0.96	0.82
23	0.81	0.76	0.96	0.84
24	0.78	0.71	0.91	0.80
25	0.81	0.73	0.89	0.81
26	0.67	0.67	0.83	0.72
27	0.69	0.66	0.96	0.77
28	0.83	0.79	0.97	0.86
29	0.74	0.69	0.95	0.79
30	**0.54**	**0.53**	**0.57**	**0.55**
Average	0.72	0.69	0.87	0.76

(b) Dataset 2

and their models involve less of discovering minute primitive patterns directly from data.

It is our hope that the presented work boosts the use of temporal models with higher-order temporal structure to exploit the advantages of the both worlds in top-down and bottom-up modeling.

Chapter VIII

DISCUSSION

8.1 Summary

The thesis statement presented in Chapter 1 can now be restated with all the terms explained in detail, and all the claims made therein defended through theoretical developments and experimental results :

> Switching linear dynamic systems with higher-order temporal structure increase the scope of the data and the temporal inference tasks that can be handled, and produce superior labeling results over the standard SLDSs.

In particular, I have presented three extensions of SLDSs which provides the following novel contributions :

1. Segmental SLDSs (S-SLDSs) produce superior labeling results by capturing the descriptive duration patterns within each LDS segment.

2. Parametric SLDSs (P-SLDSs) can model data with global variations and provides superior labeling accuracy along with the additional ability to estimate the amount of global transformation exhibited by data.

3. Hierarchical SLDSs (H-SLDSs), a generalization of standard SLDSs with hierarchic Markov chains, are able to encode temporal data which exhibits grammar-like hierarchic structure and provides the ability to label temporal data at multiple temporal granularities along with superior labeling accuracy.

Furthermore, we have described practical and tractable approximate inference algorithms for each of the above models along with the learning algorithms which optimize the model parameters based on the data under the maximum likelihood learning principle.

Finally, we have demonstrated that the SLDS models with higher-order temporal structures are practical by applying the models to two different types of data. The two applications we addressed in this dissertation are (1) honey bee dance decoding and (2) hierarchical annotation of human exercises. We demonstrated that PS-SLDSs decode honey bee dances more accurately than standard SLDSs in terms of both labeling and quantification tasks. Additionally, H-SLDSs were shown to annotate two human exercise datasets fairly accurately where each dataset was collected using different sensors.

8.2 Discussion

A number of improvements are possible to the techniques presented in this dissertation. We would describe the theoretical and practical challenges we faced in our work along with the potentially promising research directions when it is possible.

8.2.1 Labeling versus Detection

It has been observed that there are often challenges in deciding whether one should adopt the sequential labeling framework studied in this dissertation or another venue of detection framework, for the problem of automatic data interpretation. In particular, it is often the case that the particular patterns that the system designer is interested in do not densely span the whole temporal data. This phenomena appears more frequently when certain temporal patterns are over long-term temporal ranges, effectively leaving unknown segments in between. To fit the problem in the labeling framework in such occasions, we often need to relying on introducing an 'unknown' regime, as shown in Chapter 7. However, as the complexity of the data increases, the role of 'unknown' regime becomes heavier since the 'unknown' regime needs to encode larger amount of unlabeled data, rendering the learning problem for this specific regime ever more challenging. In such cases, one can attempt to adopt the detection framework since it avoids the problem of modeling gigantic 'unknown' classes. However, the limitations of detection framework is that (1) we need additional parameters for the detection window sizes and detection thresholds, and (2) more importantly, the detection accuracy for the low-level temporal patterns may be limited due to the fact that contexts are crucial to reliably identify such patterns. For example, it has been

shown in Chapter 7 that the use of hierarchy demonstrates superior ability to identify the occurrences of dumbbell exercises in comparison to the models without such hierarchies. An interesting approach would be the use of labeling approaches to identify a set of lower-level concepts, which are then fed into detection framework to identify higher-level concepts. The problem of pattern recognition and discovery on sequential categorical data has been also extensively studied, e.g., [6]. In particular, an excellent work [21] has demonstrated an in-depth study on the theoretical limit of the Markov models on the categorical sequence data. They provide insights into the relationship between the characteristics of data and the most achievable recognition rate through analytical form which is obtained through weak approximation. Indeed, the analytical analysis would provide important insights about the design issues between the labeling and detection framework depending on the characteristics of data.

8.2.2 Scalable Inference Method for H-SLDSs

As have been mentioned in Chapter 6, a more scalable inference algorithm for H-SLDSs would be needed to analyze temporal data with deep hierarchy. An interesting future direction of research for the developed H-SLDS work in this dissertation would be to develop a variant of a class of structured variational inference methods in the spirit of [37, 39]. However, the use of structured variational inference methods poses other issues such as inference scheduling across multiple Markov chains and additional levels of approximation added during the overall inference phase. An in-depth study on the various design options for such developed inference method may yield interesting insights to the research community.

8.2.3 Structure Learning Problem

The problem of learning appropriate model structure automatically from data was not deeply addressed in this dissertation. Although we have tried to explore the use of more mathematically principled measures such as AIC [3], the result was not successful even for the simple task of finding the appropriate number of LDS components to be shared within H-SLDSs, because the number of LDS components was mostly under-estimated to provide enough basis to descriptively encode the difference between different exercise categories. The learning of

deep (many layers of) hierarchy poses even further challenges. Indeed, the problem of learning deep hierarchic structure has been well-known to be a very challenging problem. Most notably, the work by Chudova and Smyth [21] has shown that even the detection problem, which is much easier than the discovery problem, can be guaranteed only limited accuracy which is bounded by Bayes' error rate. Fortunately, promising advances have been made for the problems such as deep hierarchy learning using neural networks [36] and Bayesian approaches towards the structure learning problems using hierarchical Dirichlet processes [81]. For example, a recent work has demonstrated promising results where they discover an appropriate size of the representations for SLDSs using hierarchical Dirichlet process priors [30]. A theoretical and empirical studies on the premises of all the above approaches would provide more insights into the future solutions.

REFERENCES

[1] *Proc. 17th Conf. on Uncertainty in AI (UAI)*, (Seattle, WA), August 2001.

[2] "Switching observation models for contour tracking in clutter," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 295–304, 2003.

[3] Akaike, H., "A new look at statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[4] Andrew Y Ng, M. I. J., "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[5] Ariadna Quattoni, Sybor Wang, L.-P. M. M. C. T. D., "Hidden-state conditional random fields," *PAMI*, vol. 29, no. 10, pp. 1848–1852, 2007.

[6] Bailey, T. L. and Elkan, C., "Unsupervised learning of multiple motifs in biopolymers using expectation maximization," vol. 21, pp. 51–80, October 1995.

[7] Balch, T., Dellaert, F., Feldman, A., Guillory, A., Isbell, C., Khan, Z., Stein, A., and Wilde, H., "How A.I. and multi-robot systems research will accelerate our understanding of social animal behavior," *Proceedings of IEEE*, vol. 94, pp. 1145–1463, July 2006.

[8] Balch, T., Khan, Z., and Veloso, M., "Automatically tracking and analyzing the behavior of live insect colonies," in *Proc. Autonomous Agents 2001*, (Montreal), pp. 521–528, 2001.

[9] Bar-Shalom, Y. and Fortmann, T., *Tracking and data association*. New York: Academic Press, 1988.

[10] Bar-Shalom, Y. and Li, X., *Estimation and Tracking: principles, techniques and software*. Boston, London: Artech House, 1993.

[11] Bar-Shalom, Y. and Tse, E., "Tracking in a cluttered environment with probabilistic data-association," *Automatica*, vol. 11, pp. 451–460, 1975.

[12] Bartholomew, D. J., *Latent Variable Models and Factor Analysis*. New York: Oxford University Press, 1987.

[13] Bishop, C. M., *Pattern Recognition and Machine Learning*. Springer, 2007.

[14] Brand, M. and Hertzmann, A., "Style machines," in *SIGGRAPH*, pp. 183–192, 2000.

[15] Branson, K. and Belongie, S., "Tracking multiple mouse contours (without too many samples)," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 1039–1046, 2005.

[16] Bregler, C., "Learning and Recognizing Human Dynamics in Video Sequences," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 568–574, 1997.

[17] Bui, H., Phung, D. Q., and Venkatesh, S., "Hierarchical hidden markov models with general state hierarchy," in *Proc.* 21^{th} *AAAI National Conference on AI*, (San Jose, CA), 2004.

[18] Carter, C. and Kohn, R., "Markov chain Monte Carlo in Conditionally Gaussian State Spaece Models," *Biometrika*, vol. 83, pp. 589–601, 1996.

[19] Chan, A. B. and Vasconcelos, N., "Mixtures of Dynamic Textures," in *Intl. Conf. on Computer Vision (ICCV)*, 2005.

[20] Chan, A. B. and Vasconcelos, N., "Modeling, clustering, and segmenting video with mixtures of dynamic textures," vol. 30, pp. 909–926, May 2008.

[21] Chudova, D. and Smyth, P., "Pattern discovery in sequences under a markov pattern discovery in sequences under a markov assumption," in *Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, 2002.

[22] Dempster, A., Laird, N., and Rubin, D., "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[23] Djuric, P. M. and Chun, J.-H., "An MCMC Sampling Approach to Estimation of Nonstationary Hidden Markov Models," *IEEE Trans. Signal Processing*, vol. 50, no. 5, pp. 1113–1123, 2002.

[24] Doretto, G., Chiuso, A., Wu, Y., and Soatto, S., "Dynamic Textures," *Intl. J. of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

[25] Doucet, A. and Andrieu, C., "Iterative Algorithms for State Estimation of Jump Markov Linear Systems," *IEEE Trans. Signal Processing*, vol. 49, no. 6, pp. 1216–1227, 2001.

[26] Doucet, A., Gordon, N. J., and Krishnamurthy, V., "Particle filters for state estimation of jump Markov linear systems," *IEEE Trans. Signal Processing*, vol. 49, no. 3, pp. 613–624, 2001.

[27] Duong, T. V., Bui, H. H., Phung, D. Q., and Venkatesh, S., "Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 838– 845, 2005.

[28] Ferguson, J., "Variable duration models for speech," in *Symposium on the Aplication of HMMs to Text and Speech*, pp. 143–179, 1980.

[29] Fine, S., Singer, Y., and Tishby, N., "The Hierarchical Hidden Markov Model : Analysis and Applications," *Machine learning*, vol. 32, pp. 41–62, 1998.

[30] Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S., "Nonparametric Bayesian Learning of Switching Linear Dynamic Systems," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[31] Frey, B. and Jojic, N., "Transformation-Invariant Clustering and Dimensionality Reduction Using EM," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 1–17, January 2003.

[32] Frisch, K., *The Dance Language and Orientation of Bees*. Harvard University Press, 1967.

[33] Ge, X. and Smyth, P., "Deformable Markov model templates for time-series pattern matching," in *Intl. Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 81–90, 2000.

[34] Ghahramani, Z. and Hinton, G. E., "Variational learning for switching state-space models," *Neural Computation*, vol. 12, no. 4, pp. 963–996, 1998.

[35] Ghahramani, Z. and Hinton, G., "The EM algorithm for mixtures of factor analyzers," Tech. Rep. CRG-TR-96-1, Dept. of Computer Science, University of Toronto, February 1997.

[36] Hinton, G., Osindero, S., and Teh, Y., "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[37] Howard, A. and Jebara, T., "Dynamical systems trees," in *Proc.* 20^{th} *Conf. on Uncertainty in AI (UAI)*, (Banff, Canada), pp. 260–267, July 2004.

[38] Ivanov, Y. and Bobick, A., "Recognition of visual activities and interactions by stochastic parsing," *PAMI*, vol. 22, pp. 852–872, Aug. 2000.

[39] Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L., "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, pp. 183–233, 1999.

[40] Khan, Z., Balch, T., and Dellaert, F., "A Rao-Blackwellized particle filter for EigenTracking," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 980–986, 2004.

[41] Kim, C.-J., "Dynamic linear models with Markov-switching," *Journal of Econometrics*, vol. 60, no. 1-2, pp. 1–22, 1994.

[42] Kim, S. and Smyth, P., "Segmental Hidden Markov Models with Random Effects for Waveform Modeling," *J. of Machine Learning Research*, vol. 7, pp. 945–969, October 2006.

[43] L. Xie, S.-F. Chang, A. D. and Sun, H., "Unsupervised Discovery of Multilevel Statistical Video Structures Using Hierarchical Hidden Markov Models," in *IEEE Intl. Conf. on Multimedia and Expo(ICME)*, vol. 3, pp. 29–32, 2003.

[44] Lafferty, J., McCallum, A., and Pereira, F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Intl. Conf. on Machine Learning (ICML)*, 2001.

[45] Lerner, U. and Parr, R., "Inference in hybrid networks: Theoretical limits and practical algorithms," in *Proc.* 17^{th} *Conf. on Uncertainty in AI (UAI)* [1], pp. 310–318.

[46] Lerner, U., Parr, R., Koller, D., and Biswas, G., "Bayesian fault detection and diagnosis in dynamic systems," in *Proc.* 17^{th} *AAAI National Conference on AI*, (Austin, TX), pp. 531–537, 2000.

[47] Levinson, S. E., "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language*, vol. 1, no. 1, pp. 29–45, 1990.

[48] Li, Y., Wang, T., and Shum, H.-Y., "Motion texture : A two-level statistical model for character motion synthesis," in *SIGGRAPH*, 2002.

[49] Liao, L., Fox, D., and Kautz, H., "Location-Based Activity Recognition using Relational Markov Networks," in *Intl. Joint Conf. on AI (IJCAI)*, pp. 1471–1476, 2005.

[50] Maybeck, P., *Stochastic Models, Estimation and Control*, vol. 1. New York: Academic Press, 1979.

[51] McLachlan, G. and Krishnan, T., *The EM algorithm and extensions*. Wiley series in probability and statistics, John Wiley & Sons, 1997.

[52] Minka, T. P., "Expectation propagation for approximate Bayesian inference," in *Proc.* 17^{th} *Conf. on Uncertainty in AI (UAI)* [1], pp. 362–369.

[53] Minnen, D., Starner, T., Essa, I., and C.Isbell, "Discovering characteristic actions from on-body sensor data," in *IEEE Intl. Sym. on Wearable Computers (ISWC)*, 2006.

[54] Moore, D. and Essa, I., "Recognizing multitasked activities using stochastic context-free grammar," in *Proceedings of Workshop on Models versus Exemplars in Computer Vision*, 2001.

[55] Morency, L.-P., Quattoni, A., and Darrell, T., "Latent-dynamic discriminative models for continuous gesture recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[56] Murphy, K. and Paskin, M. A., "Linear-time inference in Hierarchical HMMs," in *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[57] Neal, R. and Hinton, G., "A view of the EM algorithm that justifies incremental, sparse, and other variants," Kluwer Academic Press, 1998. Also published by MIT Press, 1999.

[58] Neuts, M. F., *Matri-geometric solutions in stochastic models*. The Johns Hopkins University Press, 1981.

[59] North, B., Blake, A., Isard, M., and Rottscher, J., "Learning and classification of complex dynamics," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 9, pp. 1016–1034, 2000.

[60] Oh, S. M., Ranganathan, A., Rehg, J., and Dellaert, F., "A Variational inference method for Switching Linear Dynamic Systems," Tech. Rep. GIT-GVU-05-16, GVU Center, College of Computing, Georgia Institute of Technology, 2005.

[61] Oh, S. M., Rehg, J. M., Balch, T., and Dellaert, F., "Data-Driven MCMC for Learning and Inference in Switching Linear Dynamic Systems," in *Proc.* 22^{nd} *AAAI National Conference on AI*, (Pittsburgh, PA), pp. 944–949, 2005.

[62] Oh, S. M., Rehg, J. M., Balch, T., and Dellaert, F., "Learning and Inference in Parametric Switching Linear Dynamic Systems," in *Intl. Conf. on Computer Vision (ICCV)*, vol. 2, pp. 1161–1168, 2005.

[63] Oh, S. M., Rehg, J. M., Balch, T., and Dellaert, F., "Learning and Inferring Motion Patterns using Parametric Segmental Switching Linear Dynamic Systems," *Intl. J. of Computer Vision*, vol. 77, pp. 103–124, May 2008.

[64] Oh, S. M., Rehg, J. M., and Dellaert, F., "Parameterized duration modeling for Switching lienar dynamic systems," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[65] Oh, S. M., Rehg, J., and Dellaert, F., "Segmental switching linear dynamic systems," Tech. Rep. GIT-CC-05-13, College of Computing, Georgia Institute of Technology, 2005.

[66] Ostendorf, M., Digalakis, V. V., and Kimball, O. A., "From HMM's to Segment models : A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[67] Pavlović, V. and Rehg, J., "Impact of Dynamic Model Learning on Classification of Human Motion," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 788–795, 2000.

[68] Pavlović, V., Rehg, J., Cham, T.-J., and Murphy, K., "A dynamic Bayesian network approach to figure tracking using learned dynamic models," in *Intl. Conf. on Computer Vision (ICCV)*, vol. 1, pp. 94–101, 1999.

[69] Pavlović, V., Rehg, J., and MacCormick, J., "Learning switching linear models of human motion," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 981–987, 2000.

[70] Rabiner, L. and Juang, B., "An introduction to hidden Markov models," in *IEEE ASSP Magazine*, 1986.

[71] Ren, L., Patrick, A., Efros, A., Hodgins, J., and Rehg, J. M., "A Data-Driven Approach to Quantifying Natural Human Motion," *ACM Trans. on Graphics, Special Issue: Proc. of 2005 SIGGRAPH Conf.*, vol. 24, pp. 1090–1097, August 2005.

[72] Rosti, A.-V. and Gales, M., "Factor analyzed hidden markov models," in *Intl. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, 2002.

[73] Rosti, A.-V. and Gales, M., "Rao-blackwellised Gibbs sampling for switching linear dynamical systems," in *Intl. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, vol. 1, pp. 809–812, 2004.

[74] Roweis, S. and Ghahramani, Z., "A Unifying Review of Linear Gaussian Models," *Neural Computation*, vol. 11, no. 2, pp. 305–345, 1999.

[75] Roweis, S. and Saul, L., "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2223–2326, December 2000.

[76] Russel, M., "A segmental HMM for speech pattern matching," in *Intl. Conf. Acoust., Speech, and Signal Proc. (ICASSP)*, pp. 499–502, 1993.

[77] S. Belongie, K. Branson, P. D. and Rabaud, V., "Monitoring Animal Behavior in the Smart Vivarium," in *International Conference on Methods and Techniques in Behavioral Research*, pp. 70–72, 2005.

[78] Schindler, G. and Dellaert, F., "A Rao-Blackwellized parts-constellation tracker," in *ICCV Workshop on Dynamical Vision; International Conference on Computer Vision*, 2005.

[79] Sha, F. and Saul, L. K., "Large margin hidden markov models for automatic speech recognition," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1249–1256, 2007.

[80] Shumway, R. and Stoffer, D., "Dynamic linear models with switching," *Journal of the American Statistical Association*, vol. 86, pp. 763–769, 1992.

[81] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M., "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, 2006.

[82] Tipping, M. and Bishop, C., "Probabilistic principal component analysis," Tech. Rep. NCRG/97/010, Neural Computing Research Group, Aston University, September, 1997.

[83] Torralba, A., Murphy, K. P., and Freeman, W. T., "Sharing features: efficient boosting procedures for multiclass object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 762–769, 2004.

[84] Truyen, T. T., Phung, D. Q., Bui, H. H., and Venkatesh, S., "Hierarchical semi-markov conditional random fields for recursive sequential data," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[85] Veeraraghavan, A., Chellappa, R., and Srinivasan, M., "Shape-and-behavior encoded tracking of bee dance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 30, pp. 463 – 476, March 2008.

[86] Vidal, R., Chiuso, A., and Soatto, S., "Observability and identifiability of jump linear systems," in *IEEE Conference on Decision and Control*, vol. 4, pp. 3614–3619, 2002.

[87] Wilson, A. D. and Bobick, A. F., "Parametric Hidden Markov Models for Gesture Recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 9, pp. 884–900, 1999.

[88] Xuan, X. and Murphy, K., "Modeling changing dependency structure in multivariate time series," in *Intl. Conf. on Machine Learning (ICML)*, 2007.

[89] Yin, J., Shen, D., Yang, Q., and Li, Z. N., "Activity recognition through goal-based segmentation," in *Nat. Conf. on Artificial Intelligence (AAAI)*, 2005.

[90] Zoeter, O. and Heskes, T., "Hierarchical visualization of time-series data using switching linear dynamical systems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 1202–1215, October 2003.

[91] Zoeter, O. and Heskes, T., "Multi-scale switching linear dynamical systems," in *Proceedings ICANN/ICONIP*, 2003.